Genetics of membranous nephropathy

Dr Sanjana Gupta

A Thesis Submitted for the Degree of Doctor of Philosophy

in Renal Genetics

University College London

Declaration

I, Sanjana Gupta, 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.

Contents Page

1. Abstract

Autoimmune membranous nephropathy (AMN) is a rare kidney disease. The genetics of AMN have been partially elucidated and confirmed the role of *phospholipase A2 receptor-1 (PLA2R1)* and HLA. The functional effect of the genetic variations is not fully understood. This thesis investigates these unexplored genetic aspects utilising a range of methodologies and unique cohorts.

Analysing genomic sequencing data of *PLA2R1* in 335 AMN patients identified 109 strongly associated variants; 9 with a very strong association, p-value $<10^{-50}$.

In a larger cohort of 1158 European AMN patients, the findings from previous GWAS were confirmed with a strong association with *HLA-DQA1, HLA-DRB1* and *PLA2R1*. No associations were found on a genome wide scale with clinical correlates of disease such as proteinuria, sex, and age.

HLA typing by imputation in 372 anti-PLA2R1 antibody positive and uniquely 32 anti-thrombospondin type-1 domain-containing 7A (THSD7A) antibody positive AMN confirmed the dominant HLA type in European AMN as HLA-DRB1*03:01 and HLA-DQA1*05:01; replicating previous studies. No statistically significant HLA type was identified for anti-THSD7A AMN.

Anti-PLA2R1 AMN has a different genetic risk than anti-THSD7A and anti-contactin AMN as determined by the genetic risk score (GRS), and this can help differentiate between them. Interestingly, 33% of dual antibody negative AMN is likely to be anti-PLA2R1 AMN.

AMN patients with a higher genetic risk have a younger age of onset. In a rare, undescribed cohort of 15 non-familial paediatric cases of AMN the GRS proved that these individuals did not have the same genetic risk factors as anti-PLA2R1 AMN.

Finally, the genetic risk of AMN in UK Biobank Europeans is 0.8%. Even though there is a high genetic risk for AMN this does not mean this proportion of individuals will develop AMN.

In conclusion, this thesis highlights important differences between antibody status groups, confirms previous GWAS findings and reports unique features about rare AMN cohorts.

2. Impact Statement

Autoimmune membranous nephropathy (AMN) is a rare kidney specific autoimmune disease with an incidence of 10 per million persons per year. Despite this it is the leading cause of nephrotic syndrome in adults and can lead to end stage kidney disease which has significant impacts on health, lifespan and quality of life.

This study examines the genetic aspects in a multi-faceted approach using bioinformatic tools and biochemical analyses and further the notions of how genetic factors do and do not contribute to disease risk for AMN.

The previous findings of two common coding variants within a European AMN cohort were replicated. These common variants may combine with the rare risk HLA-type to cause AMN - a rare disease. Pooled DNA sequencing of *PLA2R1* identified 109 significantly associated variants with AMN. The lead intronic variant was analysed for functional effect but was not replicable nor reliable. It is possible other unidentified intronic variants are contributory to disease.

This work also replicates both genome wide association studies demonstrating the two lead loci associated with AMN are *HLA-DQA1* and *PLA2R1*. For the first time it demonstrates that these loci are specific to anti-PLA2R1 antibody AMN. The risk HLA types in European cases of AMN are HLA-DRB1*03:01 and HLA-DQA1*05:01 which are part of the common European multigene haplotype. These appear to be specific to anti-PLA2R1 AMN.

The genetic risk for disease is different between the anti-THSD7A and anti-PLA2R1 AMN cases. This gives insight to the underlying differences with different antigen associations despite the fact they cause similar disease. The study was underpowered to detect a significant HLA type for the anti-THSD7A antibody group but I hypothesise this will be different. Most interestingly, about a third of the dual negative cases have a high genetic risk for anti-PLA2R1 AMN. Calculating the GRS in dual antibody negative AMN cases could identify these false negative cases. This could have a considerable impact by reducing the risk of unnecessary investigations such as whole-body CT scanning and upper and lower gastrointestinal endoscopy for patients which would have an associated cost saving.

Age and AMN are intrinsically linked to the genetics; a higher burden of risk variants in the HLA allele is associated with a younger age of onset of anti-PLA2R1 AMN. Further, in the largest dual antibody negative paediatric onset AMN the genetic risk for AMN is different to adult onset anti-PLA2R1 antibody associated AMN suggesting that paediatric AMN has a different aetiology.

This study also demonstrates in a European population that 0.8% of individuals have a high genetic risk for AMN. It would be interesting to elucidate environmental factors that contribute to disease onset. There is an expected discrepancy of those with a high genetic risk and the overall low number of individuals with disease. Identification of environmental factors may facilitate prevention by minimising exposure to prevent disease.

3. Acknowledgements

I would like to acknowledge and demonstrate my heartfelt gratitude to the following people without whom this study would not have been possible.

Firstly, and most importantly my primary supervisor Professor Robert Kleta. Without your guidance, assistance, questions on my work, support and understanding this project would not have been possible scientifically or practically. Thank you for giving me the opportunity to work within your group.

Secondly, to Dr Horia Stanescu, the most amazing and inspiring teacher that I have ever known. You motivate me to fully understand the details of whatever it is that we were talking or learning about. Your seminars were always the highlight of otherwise challenging days full of research failures! Your knowledge even after 6 years continues to astonish me. Thank you for your guidance over the years.

A special thanks to Professor Daniel Gale who was a welcome addition to my supervisors. Thank you for driving the project forward and your support in taking on both scientific and non-scientific challenges! A big thank you to Professor Detlef Bockenhauer who has always had useful insight and questions to ask of me and has provided endless support in all my writing endeavours.

Thank you to Dr Jill Norman who was always there as a listening ear and provided considerable expertise on the wet laboratory experiments and to Dr Sean Mason for providing support to the project and providing me with a studentship via UCB.

A special thanks extend to Dr Mehmet Tekman for being patient with me in my early bioinformatic endeavours. And to Dr Vaksha Patel for the guidance and supervision with the wet laboratory experiments and technical issues in addition to being a supportive listening ear over lunch for many years!

Gratitude is extended to the remainder of the bioinformatic group past and present – Mallory, Chris, Matt, Steffi, Mel, Omid & Catalin for helping me when I got stuck or just boosting my morale, listening to me moan and knowing when to take me to the pub!

Further thanks to the clinical supervision of Dr Neil Ashman in the membranous clinics and in the clinical research aspects, setting up the London Membranous Network and integrating me into the Membranous team. Thank you to Dr Stephen B. Walsh for being my educational supervisor and making sure I was keeping up to date and on track with renal training and all the support with future career decisions. Thank you also to Dr Ruth J. Pepper and Dr John Connolly for their clinical and academic support.

Heartfelt gratitude to all my membranous patients that consented to participate in my study with no benefit for themselves. Additionally, all the patients that I do not know that consented to participate across the country and Europe.

On a personal level thank you to my parents for helping me get to where I am today, without you both I would not be the person I am. A special thanks to them and to my parents-in-law for the extra babysitting duties and sacrifices you made so that I could get this work done. Thank you to my sister, Sonia, who has patiently listened to my woes and worries. Thank you to my brother, Sagar, despite the distance you have always shown your concern and support.

To my daughter, Anaya, thank you for the providing the opportunity to refocus and recall that no matter how difficult things were at work you were always ready to challenge me in new and different ways. Thank you for making me smile, laugh and cry and supplying a necessary distraction watching you grow.

Finally, and most importantly of all, my husband Rishi. Thank you for the encouragement and support and forcing me to carry on when I just wanted to give up. The whole study would not have been possible without your patience and understanding over these past 6 years. Thank you for ensuring a steady supply of treats and planning fun weekend breaks when you knew I needed them and of course for looking after Anaya countless times alone when I wasn't able.

4. Abbreviations

AMN Autoimmune membranous nephropathy
aPLA2Rab anti-PLA2R1 antibody
AUROC area under the receiver operating characteristics curve
BED browser extensible data format
bp base pairs
C3 complement factor 3
CBX3 chromobox 3
CEBPB CCAAT/enhancer binding protein beta
ChIP-seq chromatin immunoprecipitation sequencing
CIDP chronic inflammatory demyelinating polyneuropathy
CRS combined risk score
CTCF CCCTC binding factor
CTLD C-type lectin-like domains
dNTPs deoxynucleosidetriphosphates
ddNTPS di-deoxynucleosidetriphosphates
DNA deoxyribonucleic acid
eGFR estimated glomerular filtration rate
EMSA electrophoretic mobility shift assay
ENCODE Encyclopaedia of DNA elements
EP300 E1A binding protein P300
ESKD end stage kidney disease
EXT1 exostosin 1
EXT2 exostosin 2
FOS (fos proto-oncogene)
FOXA1 forkhead box A1
GATA2 GATA binding protein 2
GBM glomerular basement membrane
GRS genetic risk score
GTEx Genotype-Tissue Expression project
GWAS genome-wide association studies
hg human genome build
HTRA1 high-temperature requirement A serine peptidase 1
HWE Hardy-Weinberg equilibrium
IBD identity by descent
IDAT raw intensity files
IgG Immunoglobulin G
KASP kompetitive allele specific PCR
Kbp kilobase pair
kDa kilodalton
LR PCR long range polymerase chain reaction
MAF minor allele frequency
MAST Motif alignment and search tool
Mb megabases
MDS multidimensional scaling
MN membranous nephropathy
mRNA messenger ribonucleic acid
NCAM1 neural cell adhesion molecule 1

NELL-1 neural epidermal growth factor-like 1 protein
NEP neutral endopeptidase
NIH The National Institutes of Health
OR odds ratio
PASA polymerase chain reaction amplification of specific alleles
Pax-6 paired box protein
PCA principal components analysis
PCDH7 protocadherin 7
PCR polymerase chain reaction
PCR polymerase chain reaction
PLA2R1 phosophlipase A2 receptor 1
PVDF polyvinylidene fluoride
QC quality control
RAD21 (RAD21 cohesin complex component)
RCF relative centrifugal force
RNA ribonucleic acid
RPM revolutions per minute
rsID refSNP identifier
SD standard deviation
Sema3B semaphorin 3B
SIFT sorting intolerant from tolerant
SNV single nucleotide variant
sPLA2 secretory phospholipases
SSNS steroid sensitive nephrotic syndrome
SVS SNP & Variation Suite
T1DGC Type 1 Diabetes Genetics Consortium
TCF7L2 transcription factor 7 like 2
TFBS transcription factor binding site
THSD7A thrombospondin type-1 domain-containing 7A
Ubx ultrabithorax
UCSC University of California, Santa Cruz website's
UKBB UK Biobank
uPCR urinary protein creatinine ratio
UTR untranslated region
UV ultraviolet
V volts
VBP vitellogenin gene-binding protein
VCF variant call format
WGS whole genome sequencing
WTCC Wellcome Trust Case Control Consortium Controls
ZNF263 zinc finger protein 263

5. Introduction

5.1. Membranous Nephropathy

5.1.1. Epidemiology

The incidence of membranous nephropathy (MN) is 10 to 12 per million persons per year [1, 2]. Despite being a rare disease, it is the leading cause of nephrotic syndrome in European adults and progresses to end stage kidney disease (ESKD) in 30-40% of cases within 5 years [3-5]. Nephrotic syndrome is a tetrad of oedema, hypoalbuminaemia, proteinuria and hyperlipidaemia [6]. In the Western World the prevalence of MN in dialysis patients ranges from 0.5% in USA, 0.58% in Europe to 1.7% in Australia & New Zealand [7]. Pollution levels can also cause disparities within countries whereby pollution increases the risk on MN [8]. Further geographical differences occur due to the increased rates of infectious diseases (particularly hepatitis B) and secondary MN in less economically developed countries [9, 10].

MN is an adult onset disease with a peak age of onset between the fifth and sixth decades of life [11]. Over 80% of patients are over the age of 40 at presentation and MN is uncommon in children [11]. Males are more commonly affected than females which makes autoimmune membranous nephropathy (AMN) unlike most other autoimmune diseases [12]. (Ankylosing spondylitis, type 1 diabetes, inflammatory bowel disease and vasculitis also have a male predominance [13]). Membranous nephropathy can be divided in to two main types; primary and secondary.

5.1.2. Terminology

The terminology for MN is constantly evolving and primary MN is also known as idiopathic MN because the cause had previously remained elusive [9]. Currently, with increased understanding primary MN is termed autoimmune membranous nephropathy (AMN) because there is a kidney specific autoimmune response and the term idiopathic MN has become historical [14]. If an underlying aetiological factor or disease is found then the MN is deemed to be secondary. There are rare cases of alloimmune MN secondary to non-self sources of protein [14]. Differentiating between primary and secondary MN is difficult because clinical presentations and histological appearances on light microscopy are similar [9].

5.1.2.1. Primary membranous nephropathy

Primary MN accounts for approximately seventy percent of all cases of MN [9]. This thesis focuses on AMN and histological and clinical features are discussed from 5.1.4 onwards. Treatment for primary and secondary MN is different and so it is important to determine which an individual has.

5.1.2.2. Secondary membranous nephropathy

If an underlying aetiology is found to be causing MN then it is termed secondary MN. There can be subtle differences histologically and in the clinical course and so these should be sought for. For example, with an underlying rheumatological disease patients may have some systemic symptoms such as tenosynovitis. Histological differences can be subtle; mesangial proliferation may be present, the subepithelial deposits may be a different immunoglobulin class and some subendothelial deposits may be present [15], see 5.1.5 for details on histology. Investigations to exclude the

most common secondary causes of MN are undertaken in the diagnostic work up of a patient with nephrotic syndrome. Resolution of MN can be expected in most cases upon treatment of the underlying secondary cause [9].

The causes of secondary MN can be divided broadly in to four categories: cancer related, rheumatological disease related, infectious disease related and drug induced [14].

5.1.2.2.1. Secondary to cancer

The prevalence of cancer in MN is approximately 10% (range 5-20%). Most cases (86%) are associated with a solid organ tumour such as lung and prostate cancer. The remaining 14% is due to haematological malignancies which is a higher percentage than other solid organ tumours [16]. Screening for malignancies at the time of diagnosis is important as only 20% have a pre-existing diagnosis of cancer [16]. The remaining majority of cases are identified at diagnosis of MN, however there are a proportion that will develop a malignancy during their follow up within a median interval of 5 years [17]. This could be independent of MN and may be lifestyle and age related. In cancer there are different disease mechanisms proposed [18];

1. Antibodies generated against a tumour antigen that has a similar epitope to an endogenous podocyte antigen results in *in situ* immune complex formation.
2. Shed tumour antigens form circulating immune complexes that later become trapped in the capillary wall.
3. Immune complexes form in the subendothelial layer and then dissociate and reform in the subepithelial layer.

4. Tumour antigens can be directly deposited on the subepithelial layer and then react with circulating antibodies to stimulate an immune response.

5.1.2.2.2. Secondary to rheumatological disease

Rheumatological or systemic autoimmune diseases are more readily identifiable with the presence of alternative symptoms in addition to the nephrotic syndrome. There are a multitude of diseases that can cause MN but the most common is systemic lupus erythematosus, which has its own classification as stage V lupus nephritis for membranous nephropathy histological changes. Rarely rheumatoid arthritis can cause MN but this is more often due to drugs used for treatment (see 5.1.2.2.4) [15]. Approximately 19% of patients with urticarial vasculitis with glomerular renal involvement will have MN on their kidney biopsy [19]. MN is the most common glomerulonephritis in sarcoidosis but there can be a long latency period between manifestations of sarcoidosis and MN and vice versa [20]. In autoimmune thyroiditis anti-thyroid-peroxidase antibodies stimulate deposition of immune complexes in the glomerular basement membrane [15]. Other diseases rarely associated with MN are IgG4 related disease, Sjogren's syndrome, systemic sclerosis and ankylosing spondylitis [15]. The disease mechanism is proposed to either be the deposition of circulating immune complexes in the glomerulus [18] or circulating antibodies binding to a podocytic antigen with in situ formation of immune complexes [15].

5.1.2.2.3. Secondary to infectious diseases

Infectious causes for MN can be divided in to viral, bacterial and parasitic. Globally viral causes are the most common and are generally associated with active viral replication or chronic carriage [21]. MN is the most common extrahepatic

manifestation of hepatitis B virus infection [15]. In hepatitis B the hypothesised disease mechanism is the cationic small antigen passes through the glomerular basement membrane and deposits in the subepithelial layer which is later targeted by antibodies [15]. Hepatitis C and human immunodeficiency virus (HIV) have also been associated with MN. In most instances treatment with anti-retroviral treatment and clearance of the virus induces remission of MN [22]. Syphilis is a bacterial infection associated with MN and parasitic infections such as Schistosomiasis, malaria, filariasis and mycobacterium leprosy are rarely associated with MN [23].

5.1.2.2.4. Drug induced

Estimates of drug induced MN are from historical series and range from 7 to 14% [24, 25]. The current paradigm is that the drug or drug metabolite deposits on the glomerular basement membrane in the subepithelial layer and acts as an antigen stimulating an *in situ* immune response [26]. Gold salts, pencillamine and bucillamine are all used for the treatment of rheumatoid arthritis and are strongly associated with MN, necessitating proteinuria monitoring with their use [26]. Non-steroidal anti-inflammatory drugs were found to induce early MN in 10% of patients, and this was not associated with the duration of exposure to medication [27]. Unusually, 1% of patients taking captopril develop MN [28] although no recent case reports can be found. In most instances withdrawal of the drug is sufficient treatment for the resolution of MN without the need for immunosuppression [15].

5.1.2.3. Alloimmune membranous nephropathy

Alloimmune aetiologies of MN are rare and are caused by donor antigens triggering an immune response in the host. Examples of this include antenatal alloimmune MN,

de novo MN in kidney transplantation and graft-versus-host disease as in bone marrow transplant recipients [29-31] .

5.1.3. Historical aspects

In current times the recognition of nephrotic syndrome and diagnosis of MN is relatively straightforward with specific investigative tests. The description of oedema, kidney disease and proteinuria in unison was first described by Richard Bright in 1827 [32] but the symptoms of disease have manifested and been recognised much before that.

In the Hippocratic era, physicians were limited to their senses to diagnose disease. Oedema was visible with the eye and resulted in a 'morbid appearance' as well as causing great discomfort to the individual; this led to the understanding it was pathological [33]. The terminology for oedema was '*dropsy*', defined as 'a preter-natural collection of water in the head, breast, belly, or all over the body… The part swelled pits if you press it with your fingers' [34]. Medical students from Hippocrates' time have learnt to pinch the ankles checking for dropsy [34]. It was only after the seventeenth century that dropsy was considered a disorder of the blood and an imbalance of the excretory and absorbent functions of the body [34]. In 1784, nosologist and Physician William Cullen distinguished dropsy by the parts of the body the oedema affected, as it was widely accepted that dropsy was the disease and not the symptom of a disease. He alludes to the fact there may be an underlying cause as he states different outcomes in the traditional cure of blood-letting and that it may be dangerous in certain individuals [35]. It was in the late eighteenth century

that the paradigm changed and dropsy was thought to be due to an underlying disorder in solid viscera or an inflammatory process [34].

The link between oedema and proteinuria was not made until further developments and advances in scientific methods. The quote from Hippocrates, “when bubbles settle on the surface of the urine, it indicates a disease of the kidney and that disease will be protracted” is the first recorded evidence of proteinuria but this was largely ignored until the seventeenth century [36]. In 1664, Dekkers reports the addition of acetic acid turning heated urine to milk and forming a whey like substance [37]. This is now believed to be the first description of proteinuria [38]. Interestingly, it was widely thought by notable scientists such as Domenico Cotugno, Charles Darwin, Erasmus Darwin and Guillaume Dupuytren that the coagulable (proteinuric) urine was beneficial and was in fact a sign of the dropsy being successfully excreted from the body [39]. Chemical analysis and comparison of urine in healthy and diseased individuals led to the distinction that proteinuria was only present in disease. In 1807, the distinction of causes of dropsy were identified by comparing the urine of oedematous patients with and without liver disease; liver patients uniquely had no albumin in their urine [39]. This observation was confirmed by other scientists around the world and heating urine to test for albumin became a standard test from 1800 onwards [39]. Proteinuria and kidney disease were not linked together until later.

The first documented report of oedema and contracted kidneys from post mortem examination was made by Rufus of Ephesus (ca. A.D. 100) but this was largely forgotten despite other similar observations [38]. Richard Bright with his medical

research unit for clinical observation and extensive biochemical laboratory team related dropsy (oedema), coagulable urine (proteinuria) and kidney disease for the first time and called it ‘Bright’s disease’ [32]. Bright continued researching renal disease over his career and with Toynbee identified that the most important changes for Bright’s disease were within the glomerulus [40].

Advances in renal pathology staining led to an increased understanding of the microscopic changes occurring in the glomerulus. Bell first used the term ‘membranous glomerulonephritis’ to describe a subcategory of Ellis type II glomerulonephritis (characterised by insidious onset, proteinuria and oedema) [41]. David Jones used periodic acid-Schiff-silver methenamine stain on renal biopsy specimens to distinguish membranous glomerulonephritis as a unique morphologic appearance [42]. Development of renal pathology techniques in the 1950s with immunofluorescence and electron microscopy helped further determine unique morphological features of membranous glomerulonephritis [41]. Membranous nephropathy replaced the term membranous glomerulonephritis due to the lack of significant glomerular inflammation [43].

5.1.4. Clinical features

MN has an insidious onset and is first noticed by patients with peripheral oedema. The signs and symptoms of MN are due to the proteinuric aspect of disease and the majority of patients (80%) have resulting nephrotic syndrome [10]. Nephrotic syndrome is defined as a tetrad of signs and symptoms [44]:

- Proteinuria: urinary protein creatinine ratio (uPCR) >3.5g per 1.73m^2 of body surface area per day. It is the proteinuria that is the fundamental pathological feature that results in the other signs and symptoms of the nephrotic syndrome.
- Oedema: caused predominantly by sodium retention, noticeable in gravity dependent regions or areas where skin is thin. Oedema can be profound and can account for an additional 30% increase in body weight [45].
- Hypoalbuminaemia: for the classification of nephrotic syndrome this is <30g/L however higher levels are routinely seen despite heavy proteinuria [46]. A modern definition proposes a reduction less than the reference range (<35g/L) should be considered diagnostic [47].
- Hypercholesterolaemia: is present and can be severe with a total cholesterol >10mmol/L. The dyslipidaemia is not restricted to just cholesterol levels and include elevated serum levels of triglycerides, lipoproteins and decreased lipase activity. The increased risk of atherosclerosis and thromboembolism are linked to dyslipidaemia [48].

The remaining 20% have lower levels of proteinuria and some may be asymptomatic and proteinuria is identified on routine screening [10].

In addition to the proteinuria there are often other clinical features present. Hypertension is common amongst MN patients and a recent study found a prevalence of 45.5% with a blood pressure of ≥140/90mmHg [49]. Microscopic haematuria can be associated in MN but isolated microscopic haematuria without proteinuria is rare and occurs in <1% [10, 50]. Venous thromboembolism can also be a presenting feature or complication from MN. This is due to the nephrotic syndrome

and so is not unique to MN alone. Rates of thromboembolism have a broad range with a small study of 100 MN patients quoting a prevalence of 36% [51] whereas a larger study of 898 patients report a prevalence of 7% [52]. The single independent predictor of venous thromboembolic events is hypoalbuminaemia with a serum level <28g/L being significant [52]. Arterial thrombosis rates are increased in MN despite correction for age, hypertension, gender and smoking status [44]. Renal dysfunction may be present at diagnosis [10]. Clinical features can be variable and so the diagnosis is cemented with a renal biopsy and characteristic findings.

5.1.5. Histology

MN is defined on histopathological findings in renal tissue. It is a disease of the glomerular basement membrane (GBM) and is characterised by subepithelial immune complex deposits [53]. Periodic acid-Schiff-silver methenamine stain (silver stain) highlights the glomerular basement membrane and is therefore useful to identify the histological changes. Histological features are important to determine the diagnosis of MN. General features of tubulointerstitial disease and vascular sclerosis are associated with reduced renal survival but this is not independent of clinical features. While the stages of MN do not predict renal survival, the synchronicity of electron microscopy deposit are associated with renal progression rate [54]. There are four histological stages of MN, Figure 5.1.

5.1.5.1. Stage I

Light microscopy of stage one MN may have no visible alteration. The only noticeable difference may be a slightly thickened and prominent GBM. The silver stain highlights the GBM but not the immune complex deposits. The deposits may be

seen as pinpoint holes in the GBM. Immunofluorescence shows finely granular serum immunoglobulin G (IgG) and complement factor 3 (C3) deposits. Electron microscopy will show small scattered electron dense deposits in the GBM [3, 53].

5.1.5.2. Stage II

Light microscopy of stage two MN has an easily visible thickened GBM with silver stain that appears opaque black. The GBM matrix reacts to the immune deposit and becomes thickened and projects towards the urinary space. In early stage two the pinpoint ‘holes’ will be more prominent and as the stage progresses these holes will result in ‘spikes’. A typical ‘spike’ like appearance is caused as the immune deposits increase in size encompassing more and more of the GBM. As the immune deposits do not stain black they appear as indentations within the GBM. Immunofluorescence depicts linear granular IgG staining of the GBM. Electron microscopy depicts the subepithelial deposits which are electron dense. The GBM can be seen encircling the immune deposit. Podocyte foot process effacement is diffuse [3, 53].

5.1.5.3. Stage III

As disease advances in stage three the GBM may appear even thicker than stage two with the spikes present and more pronounced due to increasing GBM matrix reaction. The matrix reaction can envelop the deposits which can be seen as a ‘ladder’ or ‘chain like’ appearance on light microscopy and more clearly on electron microscopy. Segmental sclerosis and tubulointerstitial fibrosis may be present because of disease progression [3, 53].

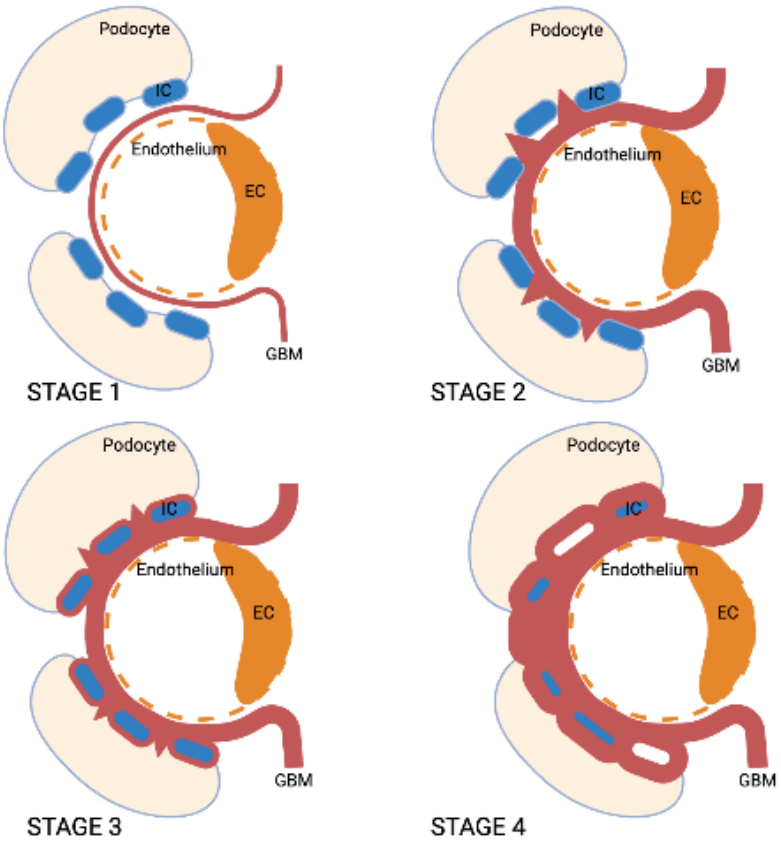


Figure 5.1: Stages of membranous nephropathy as seen on light microscopy. Adapted from Lai *et al.* [3]
Key: IC=immune complex, EC=endothelial cell, GBM=glomerular basement membrane

5.1.5.4. Stage IV

In stage four deposits are incorporated in to the GBM resulting in an irregularly thickened appearance with few spikes. Vacuoles are seen as deposits become rarefied and lose their electron density so appear irregular and electron lucent on electron microscopy. Glomerulosclerosis and tubulointerstitial fibrosis ensue [3, 53].

The stages are useful indicators of the severity of histopathological findings, yet they do not explain why the immune deposits are laid in the subepithelial layer. There are different proposed mechanisms for the pathogenesis.

5.1.6. Pathogenesis

Studies in an experimental rat model of membranous nephropathy (MN), termed Heymann's nephritis, led to the current paradigm in the pathogenesis of AMN [55]. An autoimmune response was triggered against a podocytic antigen called megalin. The circulating antibodies bound to megalin on podocytes *in situ* and formed the antigen-antibody complex [56]. This antigen-antibody complex is an electron dense immune deposit that is pathognomonic for MN and is found on the subepithelial surface of the GBM. Megalin is not found in human podocytes yet the paradigm shifted from non-renal exogenous immune complex deposition to local *in situ* immune complex formation. The mechanism was confirmed in human studies in 2002 although with an exogenous antibody [29]. The first confirmatory evidence for an autoimmune process came in 2009 by Beck *et al.* [57]. The circulating antibodies are predominantly IgG4 but the cause for their production remains uncertain [10]. The antibodies are directed against podocyte membrane expressed proteins of which a few have now been identified.

5.1.7. Human podocyte antigens

5.1.7.1. PLA2R1

The seminal discovery of the dominant human podocyte antigen in AMN was made by Beck *et al*. in 2009 [57]. Using glomerular extracts from 37 AMN individuals they

identified a 185 kilodalton (kDa) protein present in 70% of individuals but not present in 52 controls and 8 individuals with secondary MN. The glomerular extract protein was further elucidated with mass spectrometry and then tested using available antibodies and recombinant proteins. Finally utilising serum from an individual with AMN led to the discovery that the podocyte antigen was a protein called M-type phospholipase A2 receptor-1 (PLA2R1) [57].

5.1.7.1.1. PLA2R1 function

PLA2R1 is a transmembrane glycoprotein that regulates biological responses elicited by secretory phospholipases (sPLA2) [58]. There are multiple groups of sPLA2s but the one of interest in humans is group IB. PLA2R1 has dual functionality to sPLA2 – it is both inhibiting and stimulating. The stimulatory activation of PLA2R1 can induce a multitude of effects such as cell proliferation [59], cell migration [60], hormone release [61], lipid mediator production [62] and cytokine production [63]. PLA2R1 knockout mice had lower cytokine levels and were resistant to lethal lipopolysaccharide injection compared to wild type mice [63, 64]. The inhibitory action is due to the ability of PLA2R1 to internalise and degrade sPLA2 [58, 62, 65]. It is thought that this may reduce the total circulating sPLA2 and prevent heightened pro-inflammatory cytokine responses [58].

5.1.7.1.2. PLA2R1 structure

PLA2R1 is a member of the mannose receptor family [66]. The structure of PLA2R1 comprises of three parts; a large glycosylated extracellular portion that interacts with ligands, a single transmembrane region and a short cytoplasmic tail [67]. The extracellular portion is divided further in to 10 domains; an N-terminal cysteine rich

ricin domain, a fibronectin-like type II domain and eight repeated C-type lectin-like domains (CTLD), Figure 5.2. The conformation of mannose receptor family proteins is different and dependent on chemical conditions and PLA2R1 is sensitive to reduction suggesting it requires disulfide bonds for its structure [57]. The extracellular portion of PLA2R1 can either be extended or bent which affects the exposure of the epitope(s) [57].

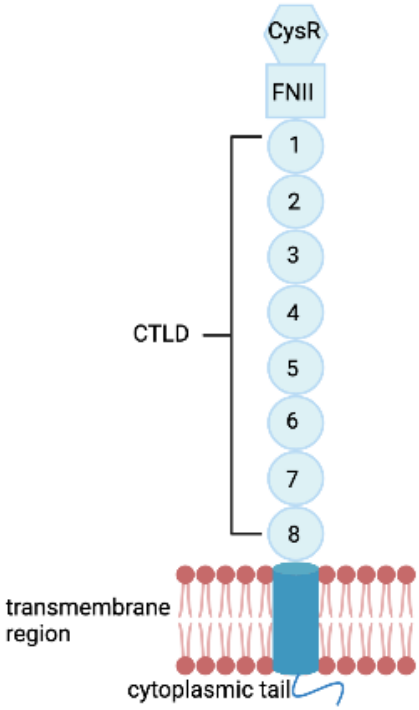


Figure 5.2: Structure of PLA2R1 (adapted from Kao *et al.*) [67]
Key: CysR=N-terminal cysteine rich ricin domain FNII=fibronectin-like type II domain CTLD=C-type lectin-like domain

The immunodominant antigenic epitope in PLA2R1 is contained within the extracellular portion of PLA2R1. It was initially identified to be contained within the complex from the N-terminal cysteine rich ricin domain to CTLD1 with the fibronectin-like type II domain containing the disulfide bonds that brings the flanking regions together for epitope formation [67]. Fresquet *et al.* further elucidated the smallest epitope in non-denaturing conditions to be the N-terminal cysteine rich ricin domain in isolation and the peptide that forms the focus of the epitope [68]. The disulfide bonds hold the three lobed structure of PLA2R1 together, which each lobe containing 12 antiparallel beta strands [68]. Other larger epitopes are N-terminal cysteine rich ricin domain to CTLD2, to CTLD3 and to CTLD8. The larger epitopes up to CTLD3 and CTLD8 are resistant to denaturing suggesting additional conformational properties that protect the epitope [68]. This is interesting as clinical studies have found that epitope spreading is associated with a worse prognosis for

AMN with lower rates of spontaneous remission and higher levels of proteinuria [69]. This suggests that the larger epitope is more stable and therefore causes more severe nephrotic syndrome.

5.1.7.1.3. PLA2R1 gene

The *PLA2R1* gene encodes PLA2R1. The gene consists of 121,867 base pairs and is found on the long arm (q) of chromosome 2 starting at position 160,797,260 and ending at 160,919,126 (human genome build 19) [70]. It is on the reverse strand and has 30 exons [70]. There are 67 orthologues across different species sets providing evidence of its importance [70]. The National Institutes of Health (NIH) Genotype-Tissue Expression project (GTEx) show with ribonucleic acid (RNA) quantification from 8555 samples of 53 tissues from 570 donors that the highest median expression of *PLA2R1* is in the thyroid with a considerable expression in the kidney cortex [70]. The kidney cortex is where the glomeruli reside in the kidney which is the site of damage in AMN. Other projects such as The Human Protein Atlas show that the second highest expression of PLA2R1 is in the kidneys [71].

The identification of the PLA2R1 epitope within the extracellular portion of PLA2R1 raised the question of variants in *PLA2R1* corresponding to the epitope. The only identified single nucleotide variants (SNVs) are in CTLD1 and in the linker between CTLD 6 and 7 [72, 73]. Amino acid substitutions in *PLA2R1* (corresponding to the SNVs in CTLD1) did not change the effect of binding to anti-PLA2R1 antibody (aPLA2Rab) compared to PLA2R1 without the substitutions [67]. This provided evidence that these SNVs in *PLA2R1* may not be responsible for the epitope presentation of PLA2R1 [67]. Further, SNVs have not been reported in the smallest

epitope of PLA2R1 (the N-terminal cysteine rich ricin domain) [68] so the significance of peptide altering variants within *PLA2R1* remain unknown.

5.1.7.1.4. Clinical associations of aPLA2Rab

Serological testing for the measurement of aPLA2Rab is commercially available and is now routine in clinical practice. It is widely accepted that the measurement of aPLA2Rab is a useful biomarker in establishing the diagnosis of AMN and the predicted specificity of aPLA2Rab measurements is 98% [74]. Despite this, renal biopsy with characteristic histopathological findings remain the gold standard [74]. Detection of PLA2R1 in the subepithelial immune deposits with immunohistochemistry or immunofluorescence depicts linear granular deposition of PLA2R1. Immunohistochemistry can also be useful to facilitate retrospective diagnosis as PLA2R1 remains detectable in paraffin embedded renal biopsy samples [75].

The clinical outcomes of AMN are highly variable irrespective of the severity of their nephrotic syndrome. A third of patients have a spontaneous remission, a third continue to have proteinuria and a third progress to ESKD [14]. Anti-PLA2R antibodies are useful to predict prognosis of AMN [76-80]. Patients predicted to progress and not spontaneously remit are treated with immunosuppressive agents. Titres of aPLA2Rab are useful in guiding treatment as rates for spontaneous remission are low with very high titres of aPLA2Rab [81, 82]. The disappearance of aPLA2Rab predate the onset of remission whereas ongoing persistence of aPLA2Rab despite immunosuppressive treatment predicts an unfavourable outcome [76, 81].

There is a wide range of reports of aPLA2Rab in serum between 52 to 86% due to the heterogeneity of studies and the method of aPLA2Rab measurement [83]. The two methods are Western blot or immunofluorescence assay test (commercially available, EUROIMMUN). The sensitivity of the immunofluorescence assay is slightly lower than Western Blot (92%) [84]. Further aPLA2Rab can be absent in serum despite PLA2R1 associated AMN [83]. Up to 24% of patients can have PLA2R1 immune complex staining in renal biopsy with undetectable serum aPLA2Rab [85]. The current hypothesis is that aPLA2Rab binds to PLA2R1 in the kidney and are only detectable when binding sites in the kidney are saturated [86]. This would explain why aPLA2Rab can be undetectable at disease onset but during follow up of the patient can be subsequently detected [86, 87]. The presence of aPLA2Rab is highly specific for a diagnosis of PLA2R1 AMN [74]. For this reason aPLA2Rab positivity alone may be considered sufficient for a diagnosis of AMN [88].

5.1.7.1.5. Anti-PLA2R1 antibody and gene interplay

Anti-PLA2R circulating antibodies are detectable in patient's serum in cases where PLA2R1 is the apparent underlying causative antigen. Injecting aPLA2Rab in mice does not induce proteinuria [89] as only 50% of human aPLA2Rab cross reacts with mice PLA2R1 [90]. Human aPLA2Rab does cross react with rabbit PLA2R1 though it is not yet known if this induces the same phenotype as AMN [90]. Antibody status is a way of subgrouping AMN patients with a well demarcated phenotype. *PLA2R1* genetic risk alleles are associated with detectable levels of the pathogenic aPLA2Rab [91]. When patients were divided into low- or high-risk *PLA2R1* genotypes, only 4% of those with the low-risk genotype had detectable aPLA2Rab

compared to 76% of those with the high-risk genotype [91]. This association was further strengthened for the detection of aPLA2Rab after combination with the low- or high-risk *HLA-DQA1* genotypes with 0% versus 73% respectively [91]. A larger study compared aPLA2Rab positive to negative patients and found the *PLA2R1* association only with aPLA2Rab positivity [92]. In aPLA2Rab negative patients compared to controls there was no association with the *PLA2R1* SNVs [92].

This is relevant as increased aPLA2Rab correlates with clinical progression of disease; with higher titres associated with ESKD at five years and lower rates of spontaneous remission [93].

However, in an Indian cohort, there was no significant association between aPLA2Rab status and *PLA2R1* SNVs [94]. Instead, there was an association with the *HLA-DQA1* risk allele and aPLA2Rab positivity [94]. This was subsequently replicated in a European cohort and the presence of the risk alleles in either a heterozygous or homozygous state in *HLA-DQA1* and *–DQB1* was significantly associated with higher circulating aPLA2Rab [93]. Neither of the two tested *PLA2R1* SNVs were associated with aPLA2Rab titres [93]. A Chinese study demonstrated *HLA-DQB1* has a strong association with aPLA2Rab positivity [95]. However these should all be interpreted with care as aPLA2Rab are occasionally found in patients without AMN [96] and the risk alleles in *PLA2R1* are present in patients with systemic lupus erythematosus, albeit with considerably lower odds ratios [97].

5.1.7.2. Thrombospondin type-1 domain-containing 7A

Thrombospondin type-1 domain-containing 7A (THDS7A) is a 210 kDa glomerular protein that was identified in December 2014 using serum from patients without aPLA2Rab [98]. It localises like PLA2R1 to the podocyte but unlike PLA2R1 it is highly enriched within the endocytic compartment, foot process and slit diaphragm [99]. THSD7A also requires non-reducing conditions for reactivity with serum, like in the case of PLA2R1 suggesting that the conformation of the epitope is vital for recognition by antibodies [98]. Like aPLA2Rab anti-THSD7A antibodies are predominantly IgG4 [98].

The structure of THDS7A is similar to PLA2R1 with an extracellular portion that has 21 domains and 1 coiled domain, a transmembrane region and a cytoplasmic tail [100]. There are multiple epitope domains within the extracellular portion but the majority of patient serum samples (87%) recognised the most terminal first and second domains [100]. Lower anti-THSD7A antibodies were detected in serum recognising smaller epitope domains and loss of epitope recognition correlated with remission of proteinuria [100].

In the initial study whereby THSD7A was first identified it was found in 9.74% of aPLA2Rab negative individuals but not in patients with other glomerular disease or healthy controls [98]. Overall THSD7A is rare, a meta-analysis of 4121 individuals found the prevalence in all AMN patients to be 3% and in aPLA2Rab negative patients this increases to 10% [101]. The rates of malignancies are higher in patients with anti-THSD7A antibodies ranging between 6 to 25% and so screening for cancer is important in these patients [101]. Anti-THSD7A and anti-PLA2R1 antibodies are

not mutually exclusive and can co-exist in approximately 1% of AMN patients [102]. A sequence of 28 amino acids within the THSD7A epitope has a similar sequence homology to the PLA2R1 epitope with autoantibody cross reactivity possible at both sites *in vitro* [103].

5.1.7.3. Exostosin

Exostosin 1 (EXT1) and exostosin 2 (EXT2) were identified by mass spectrometry of micro-dissected glomeruli from PLA2R1 negative AMN cases in June 2019 [104]. They localise to the GBM and there is a strong association with autoimmune diseases, in particular systemic lupus erythematosus [104]. EXT1 and EXT2 positivity is associated with lupus nephritis with up to 32% of positivity in these individuals [105]. There are important clinical associations with EXT1 and EXT2 patients being younger, lower serum creatinine levels, higher proteinuria and a reduced rate of decline to ESKD [105].

5.1.7.4. Neural epidermal growth factor-like 1 protein

Neural epidermal growth factor-like 1 protein (NELL-1) was identified in January 2020 with the same methodology, using mass spectrometry on glomeruli, in aPLA2Rab negative AMN patients [106]. NELL-1 was not present in PLA2R1 AMN and healthy controls [106]. Anti-NELL-1 antibodies were identified in the serum and IgG and NELL-1 co-localised in the GBM further supporting the role of NELL-1 as an antigen in AMN [106]. Histopathology is unique to NELL-1 AMN with incomplete capillary loop staining and dominant IgG1 subclass staining [107]. The prevalence of NELL-1 AMN in PLA2R1 and THSD7A negative AMN cases is 3.8% [107].

Malignancy was present in 33% of NELL-1 AMN patients which is a higher proportion compared to PLA2R1 and THSD7A AMN [107].

5.1.7.5. Semaphorin 3B

Another antigen associated with AMN was identified in glomeruli using mass spectrometry in November 2020, Semaphorin 3B (Sema3B) [108]. Sema3B was present in 4% of PLA2R1 negative AMN patients (n=11) [108]. Interestingly, 73% of these individuals were paediatric patients. The age of onset of AMN in the majority (63%) of these paediatric patients was under the age of two and the remaining 37% were teenagers between the ages of 14 to 17 [108]. In the adult onset cases the average ages was 36 years which is considerably lower than PLA2R associated MN [108]. Thus it would appear that Sema3B is associated with paediatric and young adult onset AMN [108].

5.1.7.6. Serine protease high-temperature requirement A serine peptidase 1

Serine protease high-temperature requirement A serine peptidase 1 (HTRA1) is a glomerular antigen identified using mass spectrometry and laser microdissection of glomeruli from antibody negative cases of AMN in May 2021 [109]. This antigen is expressed on the podocytes but human antibodies specific for HTRA1 have not yet been discovered in disease [109].

5.1.7.7. Protocadherin 7

Another recently identified glomerular antigen in AMN is protocadherin 7 (PCDH7) reported in May 2021 [110]. This was also identified by mass spectrometry in

microdissected glomeruli from PLA2R1, THSD7A, EXT1, EXT2, NELL1 and SEMA3B negative AMN patients [110]. PCDH7 is deposited in the GBM with a granular linear appearance on immunohistochemistry, however, the localisation to normal podocytes is not known so this may not be a true podocytic antigen [110]. IgG from PCDH7 AMN patients cross-reacted with PCDH7 [110].

5.1.7.8. Contactin-1

The most recently identified glomerular podocytic antigen, first identified in serum, is contactin-1 in October 2021 [111]. Contactin-1 is mainly expressed in neural tissues but the association with MN was first noticed and described in 1987 [112]. Prior to 2021 there were 21 cases of anti-contactin CIDP and AMN associations reported in the literature [111, 113-115]. Now, in an additional further 5 cases an association was demonstrated with anti-contactin-1 antibodies associated with chronic inflammatory demyelinating polyneuropathy (CIDP) and MN [111]. Contactin-1 was present in normal kidney glomeruli [111]. By examining the MN renal biopsies it was found that contactin-1 was co-localising with IgG4 in the GBM in only the CIDP associated MN and not aPLA2Rab AMN [111]. The time frame of AMN onset can either be concurrent with CIDP or following afterwards, so it is hard to predict when disease onset will be and therefore requires monitoring [113].

5.1.7.9. Neutral endopeptidase

Neutral endopeptidase (NEP) is located on the foot process membrane of the podocytes. This was the first human antigen discovered in 2002 from renal tissue. NEP negative mothers that are sensitised from paternal NEP in a prior pregnancy

develop anti-NEP antibodies. These antibodies cross the placenta to cause neonatal MN in future pregnancies. Adult forms of anti-NEP antibodies are rare [29].

5.1.7.10. Neural cell adhesion molecule 1

Neural cell adhesion molecule 1 (NCAM1) has been discovered to be a target antigen in membranous lupus nephritis, a form of secondary MN, and rare instances of AMN in October 2020 in microdissected glomeruli [116]. In membranous lupus nephritis the prevalence was 6.6% and in AMN 2%. Serum from patients also demonstrated the presence of circulating antibodies [116].

5.1.7.11. Debated and unidentified antigens

For the purposes of completeness, I include debated antigens here. These have been reported in the literature in small studies but their role is not confirmed nor widely accepted in the pathogenesis for AMN.

5.1.7.11.1. Alpha-enolase

Alpha-enolase co-localises with IgG4 and the membrane attack complex in immune deposits in the kidney in AMN. Autoantibodies against alpha-enolase can be detected in serum. Bruschi *et al.* report they were elevated in 25% of AMN individuals. Alpha-enolase is a cytoplasmic protein and so the clinical significance and pathogenesis remains yet to be elucidated [117]. A single further study in 2016 investigated the role of anti-alpha-enolase antibodies and found that they were highly present in both autoimmune and secondary MN [118]. Further, alpha-enolase was not identified in the subepithelial deposits and so the authors conclude that their role in pathogenesis in MN is unlikely to be directly contributory [118].

5.1.7.11.2. Aldose reductase and superoxide dismutase

Other identified antigens identified from microdissected glomeruli are aldose reductase and manganese superoxide dismutase [119]. Following this identification, IgG4 against both of these antigens was detectable in AMN patients and co-localised in electron dense podocyte immune deposits [119]. Superoxide dismutase is expressed on the plasma membrane of podocytes in AMN [119] and is present in 43% of patients at the time of biopsy [120].

5.1.7.11.3. Unidentified antigens

The number of identified antigens has increased rapidly in the last few years with advances in scientific methodology. There remains a proportion of AMN cases with unidentified antigens. The reasons for this are not fully understood. Glassock (2013) postulates this may be due to poor sensitivity of current assays, disappearance of circulating antibody prior to clinical detection of disease and in the case of PLA2R1 AMN heterogeneity of the conformational PLA2R1 epitope [121].

5.1.8. Podocyte injury

Complement activation and production of the membrane attack complex, C5b-9, is the key component causing podocyte damage in AMN. Firstly, C5b-9 induces podocytes to produce oxygen free radicals, then the podocytes produce proteases that damage the GBM. The podocytes separate and redistribute nephrin and podocin and induce a microfilament skeleton structure and disrupt the podocyte membrane. Upregulation of cyclooxygenase-2 damages the endoplasmic reticulum and the

production of transforming growth factor beta results in GBM thickening. These mechanisms all promote podocyte apoptosis [122, 123].

5.2. Genetics and AMN

In the cases of AMN it is not understood why PLA2R1 and THSD7A act as antigens and induce autoantibody formation [124, 125]. The antigens are proteins, and proteins are encoded by genes and ultimately the sequence of deoxyribonucleic acid (DNA). For this reason and the familial aggregation of AMN it was considered that genetic variants are implicated in disease.

AMN does not have traditional Mendelian inheritance, however the role of underlying genetic factors was first considered and investigated due to familial reports of disease. The role of underlying genetic factors was first confirmed by three genome-wide association studies (GWAS) and then corroborated by further targeted genotyping and association studies.

5.2.1. Familial studies

In 1984 the first case report of identical twin brothers developing MN was published [126]. A quick succession of further case reports were published, and to date nineteen families are published to have a familial form of AMN [126-140]. This indicated the possibility of a genetic component. A recent case report of monozygotic twins with the identified Asian AMN risk alleles in HLA had the same age of onset of AMN although their clinical outcomes and aPLA2Rab positivity differed [139]. Three other studies of monozygotic twins with AMN had different phenotypes with a different age of onset, clinical outcomes and progression of disease [126, 129, 130]. This suggests the environmental contribution to disease may be considerable yet is not well established. There are no comparative studies of affected identical twins

raised in different households, which would be the ideal testing condition for such an environmental study.

There is a strong male preponderance in AMN [141]. An X-linked recessive pattern of inheritance was suggested based on the clustering of disease between non-identical brothers [127, 132, 133, 135, 136, 140]. Bockenhauer *et al.* report the largest AMN family with four affected males living in different locations, excluding potentially shared causative environmental factors and providing further evidence for an underlying genetic predisposition to disease [127]. The X-linked inheritance was refuted by reports of families with apparent male-to-male transmission [131, 142] and affected members of both genders [134, 136]; thereby suggesting an autosomal genetic factor. Further support for the theory of an underlying genetic mechanism was provided by two brothers with a rare syndromic form of AMN [133]. These brothers had both AMN and deafness but no linked HLA alleles [133].

Recently, Downie *et al.* (2021) investigated the genetic component of three European ancestry families with recessive X-linked AMN, including the family initially reported by Bockenhauer *et al*. All 8 affected individuals were aPLA2Rab negative and 7 had paediatric onset of disease. An additional 18 non-affected family members were genotyped. They identified a rare 2 megabase (Mb) haplotype present in all affected individuals from all three families. There are 70 potential genes with a wide range of physiological functions within this 2Mb region. The causative or contributory gene in these families requires further investigation [140].

Familial clusters of cases of AMN are suggestive of an underlying genetic pathogenesis, but it is apparent that these do not fit a clear Mendelian inheritance pattern. The identification of a rare 2Mb haplotype in X-linked paediatric AMN is unlikely to explain the genetic component to disease in adult onset AMN. To understand the genetic contribution to disease, GWAS were conducted. Unlike linkage analysis which screens the whole genome at a family level, a GWAS screens the whole genome on a population level. Both linkage analysis and GWAS do not focus on a single candidate gene making them 'hypothesis-free' [143].

5.2.2. Genome-wide association studies

The first breakthrough in the contribution of genetic factors and AMN was with three GWASs in 2011 [73]. See 5.4.2.1 for further information on GWAS. The GWAS published in 2011 investigated three European populations with renal biopsy-proven AMN. As AMN is diagnosed from renal biopsy the phenotype is clearly demarcated avoiding difficulties in GWAS analysis such as type II diabetes mellitus which is a clinically heterogeneous disease [144]. Three GWASs were performed, 75 cases in the French study, 146 cases in the Dutch study and 335 cases in the British study. Combining the three cohorts with a total case population of 556 strengthened the association found in the individual GWAS. They identified 20 associated SNVs in *HLA-DQA1* and 13 SNVs in *PLA2R1*. The effects of the risk SNVs were examined. Even when the risk allele was in a heterozygous state the odds ratio was increased in both *HLA-DQA1* and *PLA2R1*. The strongest association was with *HLA-DQA1*, in a homozygous state of the *HLA-DQA1* risk allele the odds ratio of AMN was 20.2 and homozygosity in *PLA2R1* 4.2 [73]. This association was very robust for such a modest cohort [145], which is unusual for a GWAS [146].

5.2.3. Imputation

The SNV coverage of these initial GWAS is low compared to the coverage available with more modern technology [147], particularly of the HLA alleles [148, 149]. To further assess the strength of the SNV associations that were found in the British study an imputation study was performed. Imputation is a method to increase the statistical power of association studies and potentially identify new associated alleles, see 5.4.2.2 for further details [150, 151]. Using this method it was possible to impute and examine 8.9 million SNVs in the British cohort [147]. The strongest signals remained in *HLA-DQA1* and *PLA2R1*, and no additional loci were found as independent risk factors [147]. Imputation of classic HLA alleles was performed, with the common European DRB1*0301-DQA1*0501-DQB1*0201 haplotype showing the strongest association but providing little information beyond the lead SNV in *HLA-DQA1*. The HLA region was genotyped in more detail and this demonstrated a region of several hundred kilobase pair (kbp) in linkage disequilibrium around *HLA-DQA1* as well as other HLA class II genes [147]. Subgroup analyses were undertaken and there was no significant gender-specific genetic difference and no additional loci were found on the X chromosome [147], which is unexpected given the unusually strong male preponderance in AMN and recent findings from Downie *et al.*, but statistical power for these analyses was limited [147].

5.2.4. *PLA2R1* exon sequencing

To elucidate the specific variants within the *PLA2R1* gene, Sanger sequencing of the 30 *PLA2R1* coding exons and canonical splice sites was performed in 2013 [72]. This was an ethnically homogenous group, all 95 affected were white European and

less than half had aPLA2Rab [72]. The genetic association of the *PLA2R1* gene has been found to be greatest in those with aPLA2Rab positivity [91]. Of the variants found 6 were common and 3 in splice sites (exon-intron boundaries) but very few appeared to change protein conformation and those that did were likely private variants. One of these non-synonymous (causing amino acid alteration) common variants encodes an amino acid located within the immunodominant epitope and may have a contributory role in the pathogenesis of AMN [67, 68, 72]. One reason for the lack of exonic difference may be that the true causal variant(s) lie(s) in the regulatory region(s) of the gene. A second reason for the lack of significant results was that only 45% of the cohort had detectable aPLA2Rab thereby potentially weakening the discovery of genetic variant associations [72]. A third reason is that the combination of common variants with an environmental trigger, such as pollution (the only identified environmental trigger to date) [152], could lead to this rare disease. As the causative variant was not found in the coding region of *PLA2R1* the role of variants in the noncoding regions of *PLA2R1* were considered.

5.2.5. Genetic ancestry differences

The majority of genetic studies in AMN have utilised a candidate gene approach whereby a specific previously identified variant is genotyped [153]. The major limitation of the candidate gene approach is it can only confirm or refute previous findings and cannot detect new associations [153, 154]. Often findings are not replicated and so the reliability of these studies is questionable [153, 154].

Following the identification of the two loci in the 2011 GWAS most studies that followed were conducted in European individuals. However, it was of interest to see

if the same genetic associations were present in other ancestries and some studies set out to investigate this [12].

The same risk alleles as identified by the 2011 GWAS were present in a cohort of 114 Indian patients [73, 94]. The strongest association again was with the homozygous genotype in *HLA-DQA1* – rs2187668 but the association was also present with the *PLA2R1* SNV. The risk of AMN was 58.4 with all four risk alleles which is a strong association for a small sample size [94].

Chinese studies identified that the *PLA2R1* risk alleles increased the rate of AMN but did not influence clinical *sequelae* such as outcomes, response to treatment or reaching ESKD [155, 156]. In Western Chinese AMN patients a *PLA2R1* SNV was associated with hypertension [157]. In Chinese AMN patients the association with *HLA-DQA1* was significantly lower than in Europeans and instead *HLA-DRB1* was stronger [91, 95]. The odds ratio with the *HLA-DRB1* and *PLA2R1* risk variants was 32.4 compared to 11.1 with the *HLA-DQA1* risk alleles [91, 95, 158]. This contrasts with a study in Western Chinese individuals which found the same association with *HLA-DQA1* [159]. A study comparing AMN patients from North-western China to South China identified differences in *PLA2R1* risk alleles and allele frequency between individuals from these regions [160]. The contrasting results in these studies is due to the difficulties in the Chinese population because of genetic heterogeneity [161]. Different reference allele frequencies are found between different Chinese sub-ancestry groups [161].

Japanese AMN patients have lower rates of aPLA2Rab positivity, approximately 50-53% compared to 70% of Europeans [162, 163]. This was thought to be due to genetic differences in *PLA2R1* [162, 163]. It was interesting to find then that SNVs in *PLA2R1* were still associated with AMN in small patient cohorts [164]. In a larger cohort of 183 Japanese AMN patients four of fifteen SNVs in *PLA2R1* had an association [165]. Another study demonstrated that the lead *PLA2R1* SNV in Japanese AMN patients was not the same as the lead SNV in European, but the odds ratio was comparable [166]. In conclusion, it appears that risk variants in *PLA2R1* are present in Japanese AMN patients and so does not fully explain the difference in rates of aPLA2Rab associated AMN. A possible theory for this might be because all employed Japanese adults have an annual health screening which includes urinalysis [167, 168]. This may identify individuals with AMN before they become nephrotic and have detectable aPLA2Rab. The HLA type of Japanese AMN patients is HLA-DRB1 which is the same as other Asian ancestries [165]. *HLA-DQA1* has no association with AMN in Japanese patients [166].

In the single study of African American AMN patients, 243 individuals had six *PLA2R1* SNVs and a single *HLA-DQA1* SNV sequenced by PCR and genotyping [92]. No association was found with the *HLA-DQA1* SNV, suggesting that this may only be relevant for individuals of European and South Asian ancestry [92]. Individuals with PLA2R1 positivity on renal biopsy (n =115) had an association with the *PLA2R1* SNVs [92].

5.2.6. HLA genotyping

The HLA type is different between Asian and European patients, Table 5.1. HLA genotyping in Chinese patients identified four HLA types DRB1*13:01, DQB1*06:03, DRB1*04:05 and DQB1*03:02 associated with clinical outcomes. Having at least one of these four HLA alleles increased the risk of a decline in estimated glomerular filtration rate (eGFR) of greater than 40% during the follow up period (range 11 to 84 months) [169]. This association remained even after correction for age, gender, proteinuria, albumin, eGFR and aPLA2Rab [169]. In the UK Biobank (UKBB) (predominantly European population) DQB1*03:02 is conversely reported to be associated with increased kidney function in all subjects [170]. It is uncertain if this difference is due to the mismatch in ancestry or due to other modifiers affecting the direction of the HLA affect in AMN.

HLA Alleles	Population	Reference
DQA1*05:01	European	Stanescu *et al.* [73]
DRB1*03:01	European & East Asian	Xie *et al.* [171]
DRB1*15:01	East Asian	Xie *et al.* [171]

Table 5.1: Independent HLA types in European and East Asian ancestry in AMN

5.2.6.1. HLA genotyping of recurrent AMN post renal transplantation

A recent study of AMN recurrence in renal transplantation genotyped selected SNVs in both kidney donors and kidney recipients. Genotyping of selected SNVs known in *HLA-D* and *PLA2R1* was done in 145 donor and recipient pairs. Unexpectedly, they report that donor *HLA-D* SNVs were independently associated with the risk of

recurrent AMN in recipients. They hypothesise that the donor SNVs may alter antigen presentation or in an indirect way may render the donor kidney more susceptible to deposition of aPLA2Rab. They also produced a genetic risk score which helped provide some prediction for recurrence risk of AMN following renal transplantation. [172]

5.2.7. Genetics and immunopathology

Associations of genetic variants with AMN have been found and replicated in different studies. It is not yet fully understood how these variants contribute to disease pathology [173]. The strongest association is with the *HLA-DQA1* allele. It is hypothesised that the allele may facilitate an autoimmune response to PLA2R1 and the increased risk with the *PLA2R1* variants further exacerbates this [173]. In 2013, David Salant hypothesised there are two components that result in the interplay between the genetic variants and AMN [174]:

1. Presence of *HLA-DQA1* risk alleles that increase the risk to autoimmunity.
2. Presence of *PLA2R1* risk alleles that alters the conformation of PLA2R1 and improves its ability to become a target for autoantibodies.

Given the subsequent identification of the PLA2R1 epitope and prediction of the allelic changes with the SNVs the second part of this hypothesis seems less likely [67, 68]. It would appear instead that these variants potentially affect regulatory pathways. The role for other contributory mechanisms in autoimmune mechanisms that may contribute in AMN has been recently reviewed, see van de Logt *et al.* [83].

5.2.8. Recent GWAS

A larger multi-ethnic GWAS was published in 2020 [171]. With advances in sequencing and imputation methodology they were able to improve the genotyping coverage to 7 million SNVs. DNA from AMN patients and controls was also more readily available so the study population was 7-fold higher. The total number of AMN cases was 3782, of which 2150 were European and 1632 were East Asian descent. Ancestry matched controls totalling 9038 were genotyped with the cases. Cases and controls of both ancestries were combined to perform a multi-ethnic GWAS. The two previously identified loci in *HLA-DQA1*0501* and *PLA2R1* remained strongly associated with the European cases of AMN. In East Asian cases the HLA allele was *DRB1*1501* and in combined European and East Asian cases *DRB1*0301*. In the combined ancestry meta-analysis they identified two novel loci: *NFKB1* and *IRF4* [171].

*DRB1**0301 is one of the alleles identified in the UK Biobank as being associated with decreased kidney function [170]. The HLA types are presented as separate findings, however, it is recognised that *DRB1*0301* and *DQA1*0501* are in tight linkage with one another and form part of the second longest multigene haplotype in Europeans [175]. The full HLA type is *A*0101: C*0701: B*0801: DRB1*0301: DQA1*0501: DQB1*0201*; called the HLA A1-B8-DR3-DQ2 haplotype [176]. It is recognised that because these are in such tight linkage with each other determining which part of the haplotype will confer risk to disease is not possible [177].

NFKB1 is on chromosome 4q, the lead SNV has an odds ratio of 1.25 (1.14 in Europeans) and on conditional analysis with the lead variant the association

disappears suggesting a single haplotype. The *NFKB1* gene encodes an active DNA binding subunit of the NF-κB transcriptional complex which is known to be pro-inflammatory and is expressed in human podocytes [171].

IRF4 is on chromosome 6p, and like *NFKB1* the association with AMN is dependent on a single SNV and therefore haplotype. The odds ratio is 1.29 (1.2 in Europeans) with the lead SNV. *IRF4* negatively regulates Toll-like-receptor signalling and thereby activation of the innate immune system, however, the underlying putative mechanism could not be determined for AMN [171].

Utilising all lead SNVs a predicted heritability for AMN was 0.43 in East Asians and 0.36 in Europeans. Calculating a genetic risk score whereby the weighted sum of the relevant ethnic risk alleles the authors deduced that the disease risk in East Asians is attributable to 32% from the risk alleles and 25% in Europeans. The magnitude of this effect is very large for a GWAS [171].

5.3. Genomics

5.3.1. Overview

Genetics is the study of heredity and in particular the study of the functional unit through which is this transmitted, deoxyribonucleic acid (DNA) [178]. Genetics stemmed from the work of Gregor Mendel in 1866 as he studied discrete traits in pea plants and proved their heritability [178]. Genetics was introduced as a term by William Bateson in 1905 although the structure of DNA was not discovered until 1953 by James Watson, Francis Crick, Maurice Wilkins & Rosalind Franklin [178].

The rapid development of technology to sequence DNA has improved understanding of inheritance greatly. The structure, function and comparison of whole genomes is now possible and the study of these is termed genomics [178]. This allows study of DNA at a broader level and helps the understanding of pathophysiology and mechanism of disease [178].

5.3.2. Genetic variation

Genetic variation is the result of a mutation and is the difference between DNA sequence between individuals within a population [179]. The relationship between mutation and variation is that mutations are the source of variation [179]. A mutation is a permanent alteration to the DNA sequence [179]. *De novo* (new) mutations occur when there is an error during DNA replication that is not corrected by DNA repair enzymes. It is only once the error is copied by DNA replication and fixed in the DNA that it is considered to be a mutation. Mutations can arise from different sources and can be either deleterious (harmful), neutral (no effect on the fitness) or

advantageous (beneficial) to the organism. Deleterious mutations can reduce the reproductive function of the organism, whereas advantageous mutations may provide survival advantages. Variations can be transmitted if they occur in gametes and resultant germline cells or can be unique to an individual without transmission to offspring if they happen in somatic cells [178].

Genetic variation refers to differences either between individuals or more commonly in genomics the differences between populations [178]. Differences in gene frequencies can be estimated and help indicate the propensity for a certain trait or phenotype [178]. Variation can occur at a single nucleotide when one nucleotide is replaced by another, is deleted or a single nucleotide is inserted, SNV [180]. It is estimated that the amount of SNV variation between two individuals is approximately 0.1% [180]. Small scale insertions and deletions affect regions between 2 to 50 bp [181]. Larger areas of variation are called structural variants and this encompasses variation of DNA in a region greater than 50 bp [182]. Across the whole genome there are a greater number of smaller scale variants (SNVs up to 5,000,000) however proportionally structural variants amount to a greater proportion of the genome at 0.8% (compared to 0.08% of SNVs) [181]. The majority of variation is within the intronic, non-coding regions of DNA because there is less evolutionary pressure in these regions and alterations here do not affect the immediate fitness or survival of the organism [180, 183]. For this reason it is widely thought that they do not have an impact on the individual [180], but this is not truly known because the function of these intronic regions is not fully understood. Even if they do not have function for protein coding they may have a protective role for minimising mutations

in offspring [183]. The methodology and bioinformatic tools are more developed for the study of SNVs.

5.4. Genomic methodology

5.4.1. DNA sequencing

5.4.1.1. Targeted DNA Sequencing

There are advantages of targeted sequencing as only the target area of interest is amplified. It significantly reduces the cost of whole genome or high throughput sequencing as well as the analysis bottleneck of bioinformatics [184]. It can also be argued that it reduces ethical issues by reducing the risk of incidental findings [185]. Different methods exist for targeted sequencing and differentiating between these is important to determine which best may suit the sequencing needs.

5.4.1.1.1. Long range polymerase chain reaction

Polymerase chain reaction (PCR) is a technique to amplify and clone DNA fragments between 0.1-10 kilobases (kb) [186]. Long range PCR (LR PCR) utilises DNA polymerases that proofread the extending DNA fragment and amplify larger DNA lengths between 20 – 40 kb [187-189]. If combined with sequencing LR PCR can achieve a higher sensitivity at a lower cost and faster speed for detecting genetic variations [190].

5.4.1.1.2. Sanger sequencing

This is useful for a small number of samples and over a small genomic region. The region of interest from whole genomic DNA is amplified using DNA region specific primers and PCR [191]. The amplified DNA product is sequenced using chain termination sequencing [192]. This involves elongation of the single stranded DNA with DNA polymerase and deoxynucleosidetriphosphates (dNTPs) with additional

modified di-deoxynucleosidetriphosphates (ddNTPs) at a lower concentration to dNTPs. The ddNTPs terminate the elongation of the complementary DNA strand so multiple strands are produced of different lengths. The different lengths of DNA fragments are separated using electrophoresis and the specific base of ddNTP (tagged with a fluorophore) can be identified for each fragment using ultraviolet (UV) light [192].

5.4.1.1.2.1. DNA sample pooling

DNA sample pooling consists of pooling together DNA samples from different individuals for genotyping. This can be either tagged, whereby the individual DNA samples are provided with a unique identifier or barcode (termed multiplexing). Or it can be untagged where the DNA amplicons from the different individuals are mixed together in equimolar amounts and the individual data is not traceable but instead a proportion of sequence is provided such that allele frequencies can be derived from this information. The benefits of untagged DNA pooling is that individuals are not identifiable, so this method has been used in HIV screening [193]. Allele frequency can be measured from pooled samples rather than individual samples increasing the overall efficiency of pooling [193]. For rare mutations pooling DNA samples can reduce the number of tests that need to be performed by 50-80% depending on the test specificity [194]. A disadvantage of DNA pooling for rare mutations is the limitation to detect a mutation given that the mutation is present in a member of the DNA pool [194]. Another disadvantage is that haplotype frequency information is unavailable with pooling of the data [194].

5.4.1.1.3. KASP genotyping

Kompetitive allele specific PCR (KASP) is a method that was first described in 1989 (then known as polymerase chain reaction amplification of specific alleles, PASA) [195]. KASP genotyping differentiates from PCR as it only amplifies regions with the alleles of interest. Allele specific primers are designed to bind to the complementary region with the 3' primer tail finishing at the SNV of interest. A reverse primer extends the 5' tail of the forward primer and incorporates this tail into the DNA template for future PCR. Cycles of PCR are undertaken to amplify this product, then fluorescently labelled oligonucleotides are added. They bind to the 5' end of the original forward DNA template and generate a fluorescent signal, Figure 5.3. Both alleles have a different fluorescent dye – either FAM or HEX and so the combination of colours identifies homo or heterozygous alleles [196].

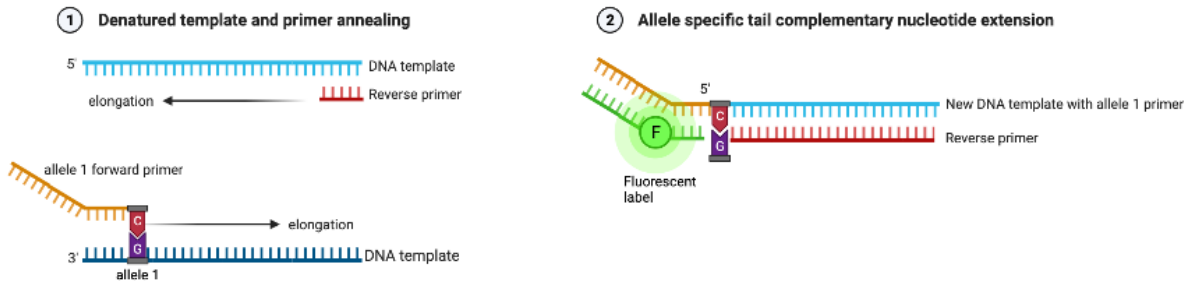


Figure 5.3: Schematic of KASP allele specific PCR, adapted from He *et al.* [196].

5.4.1.1.4. Amplicon based target enrichment

Different products available on the market have slightly different methods of amplicon based DNA target enrichment. DNA is first fragmented, which is done either with restriction enzymes or PCR primers. The region of interest is labelled with an adaptor and amplified with PCR. This is then sequenced using high throughput

sequencing (see 5.4.1.3). Tiling can be performed to increase the sensitivity and reliability of the data by using overlapping segments of sequence [197].

5.4.1.1.5. Hybrid Capture Sequencing

Different to amplicon based enrichment is DNA random fragmentation using sonication. Synthetic oligonucleotides specific to the region of interest are then added in solution and captured with streptavidin beads. Adaptors are ligated and only the captured regions are then amplified with PCR and sequenced [197].

5.4.1.2. DNA Microarray

DNA microarray is a gene chip that carries multiple probes with known DNA sequences for pre-specified SNVs. SNVs are chosen strategically to either cover the whole genome at areas of known variation or in targeted genes. DNA is amplified, fragmented, and immobilised by hybridisation to the solid substrate DNA microarray or bead chip. A single nucleotide complementary extension of the sample DNA is made on the microarray probe. DNA fragments of non-complementary binding are washed away. The nucleotide is fluorescently labelled and the bead chip is read via a laser confocal scanning machine and the fluorescent signal intensities are measured and interpreted [198, 199].

5.4.1.2.1. Raw intensity files

The raw data from an Illumina microarray is outputted into a binary format file called a raw intensity data file or IDAT [200]. There are two IDAT files; one contains data on the intensity of the colour for the red fluorescence data and the other for the green fluorescence data [201]. There are four fields in the IDAT file; the identification of

each bead on the microarray, the mean and the standard deviation of their intensities and the number of beads [201]. There is some metadata including data the microarray was scanned, software versions and the type of microarray used [201].

This file is combined with a DMAP file which is a map file for the location of the probes on the microarray which can be downloaded from the Illumina support site [200]. With the combined data of the bead locations on the microarray from the DMAP file and the quantification of the signal associated with each bead from the IDAT file these two files are combined to create a genotype call file [200].

5.4.1.3. High throughput sequencing

High throughput sequencing can sequence millions of fragments of DNA in a parallel fashion [202]. There are many different platforms for high throughput sequencing available, see [203] for a review on recent and frequently used technologies. They follow a similar principle of DNA preparation, amplification, followed by parallel sequencing [203]. The most widely used is Illumina sequencing.

In Illumina sequencing, DNA is fragmented randomly into shorter fragments and DNA adapters are attached to both ends of the DNA fragments [202]. The DNA adapters contain a complementary sequence that allows the DNA fragments to bind to the flow cell [202]. By attaching a unique identifier during the adapter ligation for each individual multiplexing can be done [202]. The DNA libraries are loaded on to a flow cell and the clusters of DNA fragments are amplified resulting in millions of copies of single stranded DNA [202]. Fluorescently labelled free floating nucleotides with reversible terminators are added to the flow cell one at a time with DNA

polymerase [202]. Complementary strands of the clusters of DNA can be created this way and if the fluorescent nucleotide in incorporated into the growing DNA strand it fluoresces [202]. The fluorescence can be detected identifying the location and the nucleotide that has been attached [202]. Through this mechanism the sequence is read contiguously of millions of clusters. Each nucleotide is sequenced over thirty times (read depth) to reduce the rate of error [204]. Once the whole sequence has been read it is aligned and overlapped to a reference genome and differences are compared to produce a variant call file [205]. It should be noted that repetitive regions are difficult to sequence as the short reads cannot be aligned to the reference genome. The schemata for Illumina sequencing is shown in Figure 5.4.

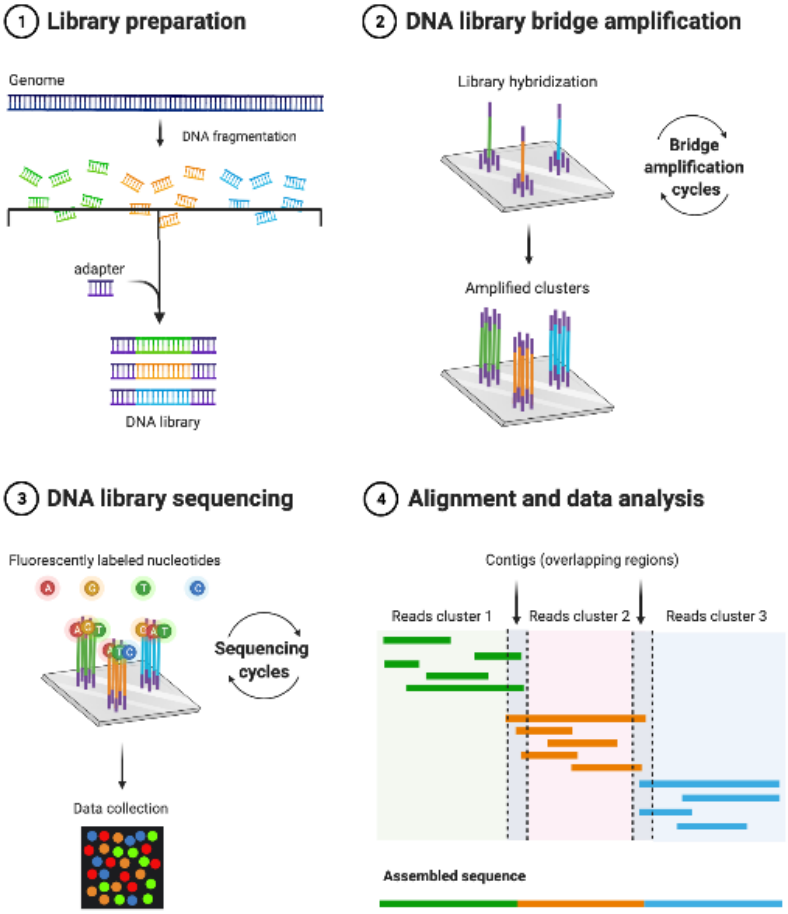


Figure 5.4: Diagrammatic representation of the Illumina high throughput sequencing schemata in a step-by-step manner. Adapted from Ansorge *et al.* [202]

5.4.2. Statistical and bioinformatic analysis

5.4.2.1. Genome wide association study

A GWAS is a statistical tool to study differences distributed over the whole genome between two populations. One population has the phenotype of interest and this is compared to a population without the phenotype. GWASs screen the whole genome and do not focus on a single candidate gene making them 'hypothesis-free' [143]. This makes them ideal for scenarios where disease pathophysiology is not fully understood, as in AMN [143].

GWASs hypothesise that the phenotype or disease is caused by variations in a subset of alleles. GWAS examine all twenty-two autosomal chromosomes for common SNVs and compare cases to controls in humans. These common SNVs can help identify the chromosomal location of interest associated with disease but are not causative [143].

The underlying basis for a GWAS is that the allele frequency at every single SNV is compared between cases and controls [206]. Allele frequencies between different ancestries vary at each SNV so cases and controls need to be ancestrally matched to prevent identification of false positives [207]. Once confounders are excluded and a difference between the allele frequency of cases and controls remains it suggests a relation between the allele or more often a nearby variant to the susceptibility of disease [206]. There can be two outcomes, either a direct association whereby the genotyped SNV is functional and has a direct association with the phenotype. This is unusual; more commonly the identified SNV is not causative and is in linkage disequilibrium with another locus and acts as a tagging marker if the other locus has not been sequenced, this is called an indirect association [208]. Data on linkage disequilibrium can be used to help identify these associations. Linkage disequilibrium is the non-random association of alleles at two or more loci and results in a higher frequency of association of the different alleles compared to if they were independent [209].

5.4.2.1.1. Epistasis

Epistasis is non-independence of effect; this can be of any genetic unit, either loci or the gene and the environment [210]. Traditionally the interaction between genes has been studied as it is easier to quantify and determine [211]. GWAS analyse SNVs and so in this context epistasis means non-independence of SNVs. It is thought that a large proportion of missing heritability is due to genetic interaction and so epistasis should be examined in a GWAS [212]. The most common method is a pairwise interaction where every SNV is analysed against all the others for non-independence [212]. The limitation with this approach is that this requires considerable computational processing power so this is normally overcome by p-value filters and limiting analysis to certain selected SNVs [212].

5.4.2.2. Whole genome imputation

This technique is based on knowledge about stretches of shared haplotype to provide information and predict the untyped alleles [213]. Imputation takes advantage of haplotype composition to match known SNVs to other SNVs that are in linkage disequilibrium with them.

Imputation uses stretches of shared haplotype to estimate the missing genotype that is present [213]. Genotype imputation was first developed in related individuals because large regions of shared haplotype were easily identifiable. The first report of *in silico* imputation from 6,500 to 53 million genotypes in related individuals was presented in 2006 [214]. In unrelated individuals the underlying principle is the same; the known haplotype is utilised to estimate the unknown genotype in another individual [213]. Because the two individuals are not related the stretches of shared

haplotype are a lot smaller [213]. The missing genotype for each study sample can this way be estimated by copying alleles from the reference haplotype (formed from many individuals) [213]. As this is an estimate the tools provide probabilities for the certainty of correct imputation at each allelic position [213].

Imputation accuracy is a concern as the genotypes are being estimated and then the association with a phenotype is predicted utilising this data. The probability of correctly imputing depends on the number of individuals in the reference panel, the density of genotyped data, the strength of linkage disequilibrium at each position and vitally the matched ancestry of the reference and case individuals [215]. Imputation accuracy is determined mathematically by the computational tools. These calculate the squared correlation between the number of minor alleles (allele dosage) of the most likely imputed genotype and the allele dosage of the true genotype (i.e. observed and expected) [216]. The allelic dosage can be filtered out and added as co-variate in GWAS analysis.

Different imputational tools exist, broadly the two main categories are either computationally intensive or more computationally efficient tools [213]. Broadly speaking computationally intensive tools consider all observed genotypes when imputing whereas the more efficient methods focus on a smaller number of nearby genotypes to the desired imputed locus

5.4.2.3. Human leucocyte antigen analysis

The human leucocyte antigen (HLA) is a group of more than 200 protein coding genes [217]. HLA gene products play a key role in antigen presentation and immune

signalling [217]. No two individuals have exactly the same HLA (apart from monozygotic twins) and the genes are highly polymorphic [218]. HLA analysis can be either at the gene level with molecular typing or at the protein level (antibodies) by serological typing [219]. Direct molecular typing of HLA alleles is often limited in large cohorts because of time constraints, labour intensive methodology, expense, reduced allele resolution and reduced HLA gene coverage [220, 221]. For this reason, HLA alleles are often indirectly imputed from SNV data using population specific HLA reference panels. The distribution and frequency of HLA alleles are highly variable across different ancestral groups which results in differences in HLA risk alleles in different ancestry populations; as seen in AMN [171, 222].

The underlying principle of HLA imputation is the same as whole genome imputation. However, due to the complex linkage disequilibrium in HLA the pattern recognition algorithms need to be adjusted [223]. A limiting factor for imputation is access to relevant ancestry haplotype data [223]. HLA imputation accuracy can improve by up to 30% by using a more closely matched reference panel [224]. The consensus is that the reference panel and sample size is more important than the SNV density for high imputation accuracy [225]. Despite high imputation accuracy the accuracy for rare alleles is lower [226]. Newer techniques using deep machine learning are now available that can impute rarer alleles, facilitate trans-ethnic analysis and all with faster computational processing times [223].

HLA association testing follows the same principles as a GWAS whereby the allele frequencies between cases and controls are compared and the statistical likelihood of an association is determined at each locus.

5.4.2.4. Genetic risk score

The genetic risk of a phenotype can be determined by the combined set of risk variants to identify individuals at increased risk for the phenotype. The genetic risk score (GRS) or polygenic risk score is a weighted sum of the number of risk alleles an individual carries [227]. GRS can predict disease status in research based case-control studies, population based cohort studies and in electronic health record based studies [228-232]. The clinical utility of GRS is established in multiple disease such as breast cancer, prostate cancer and coronary artery disease [233].

The GRS is calculated from a set of independent risk variants associated with a particular disease or phenotype based on evidence from a GWAS. For each individual, the number of risk alleles at each variant is summed (0,1,2) and is weighted by its effect size (the natural logarithm of the odds ratio for binary traits) [227]. This produces a single score for each individual's cumulative genetic risk for the phenotype of interest [227]. This can be summarised mathematically as in Equation 5.1. The summation of the scores assumes that each SNV has an additive independent risk and so independent SNVs need to be used for the calculation. If there is epistasis within the SNVs then a different method will need to be used [227]. While the GRS model may seem simplistic, a large meta-analysis from twin studies found the majority (69%) of traits are consistent with a simple additive genetic effect [234].

$$GRS = \sum \frac{Number\ of\ risk\ alleles_{at\ each\ SNV} \times ln(odds\ ratio)_{at\ that\ SNV}}{Number\ of\ risk\ alleles}$$

Equation 5.1*:* Equation for genetic risk score (GRS). The sum of all risk SNVs is calculated by calculating the cumulative risk for each SNV.

The biggest limitation of the GRS is its applicability to different ancestries. The ancestry of the population needs to be matched to the ancestry of the GWAS population and at present this is overrepresented with European studies. GRS are unable to factor in gene and environment interactions that may exist. For example an individual may have an increased genetic risk for alcohol dependence but if they are never exposed to alcohol then they will not develop the condition [227]. The GRS is simple but it can be limited in capturing the full genetic loading for a disease because an incomplete list of SNVs is used (compared to whole genome sequencing) and therefore the causal variants themselves are not identified and the calculated odds ratios and scores can also be imprecise [227].

GRS can be useful for an individual's genetic risk stratification for a particular disease. The individual must match the characteristics of the original research study used to estimate the effect size [227]. It can be useful if screening programmes or lifestyle modification or preventative treatment can be initiated [227]. Screening programmes for cancer have considerable economic cost associated; for example an individual with a high GRS for breast cancer will need further mammograms which will likely cause unnecessary stress and anxiety [227]. It is also important to remember that genetic risk alone does not guarantee certainty of developing disease [227]. A clinical research trial is underway to assess the impact of GRS to estimate

the lifetime risk of breast cancer in patients with negative multi gene panel testing and no family history [235].

5.4.3. Functional genetic analysis

5.4.3.1. Computational approaches

Sequencing genome wide and the analysis produces many SNVs that are statistically associated with the phenotype of interest. Determining which of these SNVs is important or relevant requires further filtering. Functional assays and research are time consuming and expensive, so it is important to filter down and aim to predict functionality instead of investigating millions of variants. The process of identifying functional variants can be divided in to two different but not mutually exclusive steps [236]. The first is mapping SNVs to previously identified functional genomic features, identifying their consequences by annotation and comparison to known variants [236]. A summary of most of these data can be found on the University of California, Santa Cruz (UCSC) genome browser [237] and the Ensembl variant effect predictor [238]. Protein coding and regulatory regions such as transcription factor binding sites (TFBSs) are of particular interest. There is also value in determining regions that are highly conserved as these often contain regulatory DNA with transcription factors and are vital for development [239].

The second method uses computational tools to predict the nature and magnitude of the functional impact of the SNV [236]. Predictions can be made about the SNVs potential effect on either protein coding or regulatory regions [236]. SNVs affecting protein coding sequence use evolutionary information, secondary and tertiary structural features and properties and locations of the amino acids to determine

functionality [236]. Knowledge of features such as the size, polarity, surface accessibility, hydrogen bonding and conservation status are used [236]. The computational tools distinguish between these and produce a score for the theoretical model to help facilitate interpretation of the likelihood impact of the SNV [236]. Regulatory variants are determined through statistical approaches of the TFBS motifs based on pre-existing experimental results from DNA-protein binding experiments [236]. The ENCODE (Encyclopaedia of DNA elements) project catalogues large experimental data sets across several different tissues [236]. It is not known how stronger or weaker binding at TFBSs will affect the transcription factor function [236].

5.4.3.2. Biochemical approaches

There are different biochemical approaches to investigate functionality of genetic associations. Protein-DNA interactions are the most investigated interactions and in particular transcription factors to their binding site within DNA. Currently, the most popular method to investigate transcription factor binding to DNA sequences is chromatin immunoprecipitation sequencing (ChIP-seq) [240]. The main limitations of ChIP-seq are cost and the availability of specific antibodies for the protein of interest [240]. There are alternative methods to investigate protein-DNA interactions; electrophoretic mobility shift assay, DNA pull-down assays, microplate capture assays, and reporter assays [241].

5.4.3.2.1. Electrophoretic mobility shift assay

This method is used to study the binding of protein to its protein binding site, including binding to DNA [242]. The reaction is simulated using recombinant protein

and synthetic oligonucleotides. Utilising electrophoresis protein-DNA complexes can be separated from free protein and DNA in a gel due to differences in molecular mass [242, 243]. The bound protein DNA complex will be larger and hence ‘shift’ upwards compared to the control. There were many reasons why I chose this method to investigate the protein-DNA interactions. The most important was that our laboratory was already set up to perform this method and it did not require purchasing of any additional equipment. It is relatively basic and easy to perform while being robust to a variety of conditions [244]. Further because of its sensitivity it is possible to use low concentrations and small sample volumes in addition to the ability to use recombinant proteins which was important to save both time and cost [244].

6. Methods

6.1. PLA2R1 intronic variant analysis

6.1.1. Computational analysis: identification of variants of interest

6.1.1.1. Patient cohort

Blood samples for DNA were collected by the Medical Research Council and Kidney Research UK National DNA bank for glomerulonephritis. The diagnosis of AMN was assessed and confirmed by the UK Membranous Nephropathy Consortium. This identified 335 self-reported European ancestry patients with kidney biopsy-proven AMN.

6.1.1.2. Ethics

The study was approved by the Oxford multicentre research ethics committee, Oxford UK. The study was conducted to the Declaration of Helsinki principles. Informed written consent was collected prior to blood sample donation.

6.1.1.3. DNA preparation

DNA was extracted from blood samples, one sample had a low yield of DNA so was excluded giving a total of 334 AMN patients. Extracted DNA was sent to deCODE (Reykjavik, Iceland) for LR PCR and pooled DNA sequencing. PCR primers were designed to cover the *PLA2R1* gene by deCODE. The target gene locus spanned 160,506,258 – 160,627,367 base pairs (bp) on chromosome 2: primers were designed to cover an extended region of chr 2:160,503,502 – 160,643,828 (140 kbp)

(GRCh36/hg18). The primer pairs were designed such that there were overlapping segments of DNA. Table 6.1 demonstrates the regions covered by the primer pairs and the regions that failed are shown on *PLA2R1* in Figure 6.1. The coverage by LR PCR was 93.6%.

Chromosome position	Size (base pairs)
160502964 - 160512722	9759
160512264 - 160517088	4825
160516580 - 160522040	5461
160521263 - 160531420	10158
160529427 - 160539418	9992
160538868 - 160548896	10029
160548447 - 160558110	9664
160557692 - 160565965	8274
160565309 - 160575275	9967
160574929 - 160584896	9968
160584026 - 160593963	9938
160593383 - 160602848	9466
160602199 - 160612142	9944
160611440 - 160621422	9983
160620818 - 160629391	8574
160628728 - 160638433	9706
160636183 - 160641913	5731
160641227 - 160647248	6022

Table 6.1: Long range PCR primer pair regions with the chromosomal position on chromosome 2 and the correlating base pair size of the region target.

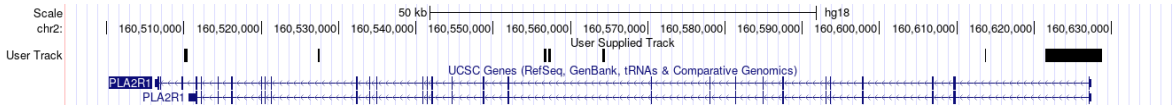


Figure 6.1: Position of missing regions not covered due to PCR primer failure relative to *PLA2R1* gene introns and exons.

6.1.1.4. DNA sequencing and alignment

An Illumina 1GA genome analyser (Illumina, San Diego, California, USA) with a paired end read length of 35 bp was used to sequence DNA. The average read depth across the entirety of the *PLA2R1* gene was 2044-fold, with a maximum coverage of 31,948 reads. The global alignment algorithm ELAND was used to align

the sequence data to the human genome build (hg) 18 [245]. There were missing regions amounting to 9.1kbp (6.4% of the total gene sequence), which were scattered throughout *PLA2R1* and are visualised in Figure 6.1. deCODE use their own in-house software for quality control and variant analysis.

6.1.1.5. Quality control

Quality control was done by deCODE within three main domains. Firstly, potential PCR bias was accounted for by excluding variants from strands of the same length exclusively carrying the minor allele. Secondly, minor allele variants close to strand endings were discarded. Thirdly, each SNV is reported with a 'SNV score', this is a logarithmic likelihood ratio test score that has been compared to a chi-square table. The higher the SNV score the more likely the SNV is real, and only scores greater than 10 were included in the report, with most real SNVs having very high scores of several hundred to thousands.

6.1.1.6. Converting case data

Patient data was provided under hg18, however, control data used the relatively newer hg19. This difference meant that the genomic positional data was not directly comparable as there was the potential of a mismatch between novel and discarded variants between the builds.

The case variant call file (VCF) file (see 6.2.1.3) was converted to the UCSC browser extensible data (BED) format [237]. The BED format of the SNV data was converted using the liftOver utility provided by UCSC [163]. The resulting BED file was converted back to a column file with updated positions. Concordance between the

variants in the different genome builds were compared and there was an excellent 1-to-1 mapping of all nucleotides between the hg18 to hg19 builds.

6.1.1.7. Control dataset

Control genotypes and allele frequencies were extracted from the European subset of individuals within 1000 Genomes data (Phase3 2013/05/02 release, Build37/hg19). Raw data in the form of a gzip-compressed VCF file was downloaded [246]. The 1000 Genomes Project specifies two pedigree files for their sample data, one for those genotyped and another for those sequenced.

The genotyping population file had 3500 individuals and quoted 669 of these as being of European (EUR) origin. However, when this was manually checked against the headers of the raw data there was a discrepancy of 166 individuals, where only 503 matched the EUR identification (ID). The whole genome sequencing (WGS) population file from the 1000 Genome Project had 2504 individuals. 503 of European origin and these all matched the EUR ID in the Phase 3 data. For this reason, the WGS population file for the control data was used.

The European group consist of a subset of individuals from Utah, America (with Northern and Western European ancestry); Tuscany, Italy; Finland; England; Scotland; and the Iberian Peninsula, Spain [247].

6.1.1.7.1. Determining allele frequencies

The 1000 Genomes Project chromosome 2 VCF file has pre-computed minor and major allele frequencies for each ethnic super group. The allele frequencies for the 503 European individuals were extracted.

6.1.1.7.2. Determining reference and alternate alleles

The FASTA data for chromosome 2 (hg19) was accessed via the UCSC website [237]. Using the chromosomal position of the 7,081,601 variants from the 1000 Genomes chromosome 2 VCF file the reference allele was extracted from the FASTA data. The quoted reference allele in the VCF file matched 100% to the FASTA file allele.

To assess the characteristics of the control data, the quoted reference allele and calculated major allele were compared over all 1000 Genomes variants. For chromosome 2, 97.1% of reference and alternative alleles mapped to the correct corresponding major and minor allele respectively. A total of 205,223 variants (2.9%) were mapped incorrectly such that the reference allele was the minor allele or the alternative allele was the major allele. In the *PLA2R1* region alone this accounted for 171 out of 3316 variants being in discord (5.2%). These were altered to match the case dataset allele with the relevant allele frequencies irrespective of the designated label.

6.1.1.8. Intersecting variant datasets

The variant datasets were compared and aligned to ascertain differences in allele frequencies within the case data and the control data. The difference in allele

frequencies was subsequently used to generate contingency tables in the undertaking of a chi-squared test to ascertain statistically significant alleles.

From the pooled sequence data, deCODE identified 962 good quality variants (set 1), 141 of medium quality (set 2) and 1200 of poor quality (set 3). There were 3316 variants in the control file within the *PLA2R1* gene. The variants from the case data were compared using the chromosomal position in the correlating control population data file. Set 1 (good quality data) had 495 variants that were present in the control data, set 2 had 8 variants and set 3 had 59 variants. As I was interested in the high-quality sequence reads, I decided to only analyse set 1.

All case variant data was bi-allelic. If the major allele in the case variant file was present in the reference allele in the controls this was a successful intersection. The control data had a few complications that needed resolving prior to full and successful intersection. The first of these was that the control variant data did not contain the variants that were present in the cases. It was important not to lose this information as it may be indicative of novel causative variants. If the control data did not have the case alternate allele, then “simulated” alleles were created to match this, see Figure 6.2. The reference major allele (from the FASTA file) was taken to create the control major allele with an allele frequency of 1 and the alternative allele with an allele frequency of 0.

Another problem with the control data was that some positions had multiple alternate alleles. This was initially resolved by matching the allele present in the variant data to the allele in the control data with the appropriate allele frequency. If the case variant

allele was matched with the allele from the control data this was called a “matching” allele. If on the other hand the allele was not present at all in the control multiple alleles this was resolved by creating an allele of the same nucleotide with an allele frequency of 0 and was termed a “forced” control allele. By reconciling the multi-allelic control data, the overall allele frequencies did not always equal 1 which potentially could create difficulties with chi-squared testing and in plotting the data for visualisation (LocusZoom). A summary of the data that was reconciled in the original attempted ways is shown, Table 6.2. As this represented few variants (13 in total), a decision was made to exclude these variants to facilitate simpler analysis.

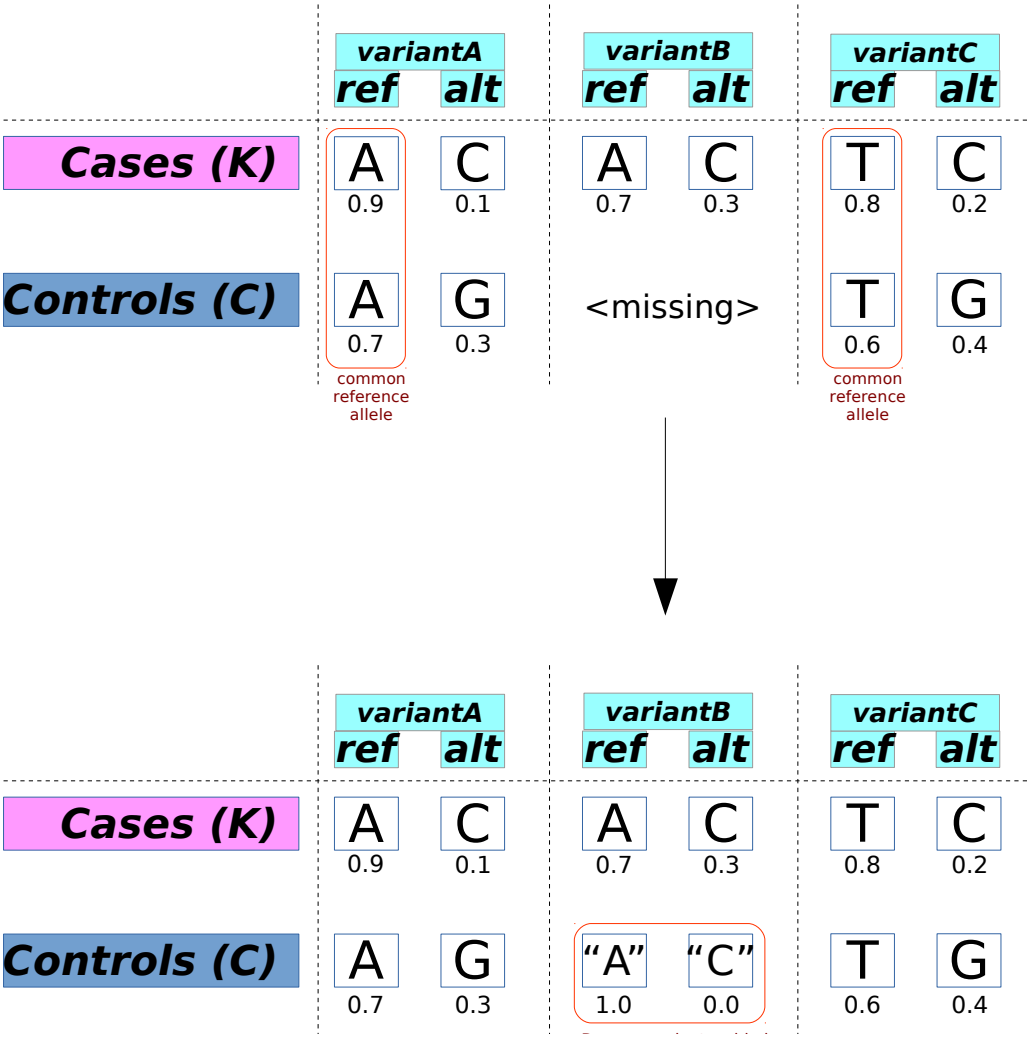


Figure 6.2: Schemata of how the simulated controls were created for variants not previously identified in the control data.

No of matching case and control alleles	No of case variants requiring a forced bi-allelic control	No of case variants requiring a control simulated variant	No of case variants matched to multi-allelic control, corrected by matching	No of case variants not matched to multi-allelic controls, corrected by forced allele	Total no of case allele variants
482	6	473	6	1	962

Table 6.2: Summary of case variant data with the matching control allele and the method used to create a matched control allele.

6.1.1.9. Chi-squared testing

6.1.1.9.1. Converting allele frequencies to absolute allele values

After obtaining a bi-allelic set of data between cases and controls for comparison a chi-squared test to compare the different allele frequencies was necessary. As chi-squared uses integer values, the allele frequencies needed to be converted to absolute allele values. To do this the allele frequency was represented as an integer. As an example: allele frequency for allele A =0.2 and allele C =0.8. If there are 10 samples then there are a total of 20 alleles (10 samples, each which have 2 alleles). Using the allele frequency for allele A the absolute allele value is 0.2 x 20 =4 alleles and then allele C is 0.8 x 20 =16 alleles. This was done for each variant and each allele.

6.1.1.9.2. Calculating observed and expected variant values

Observed values were calculated for both alleles at each chromosomal variant position for both the control data and the case data. The data was calculated with the total number of alleles in the case dataset (668) and the control dataset (1006). The observed values were calculated by Equation 6.1.

Case reference allele = $668 \times allele\ frequency$

Case alternate allele = $668 \times allele\ frequency$

Control reference allele = $1006 \times allele\ frequency$

Control alternate allele = $1006 \times allele\ frequency$

Equation 6.1: Equation used to calculate the observed values for the Chi-squared test

The expected values were the overall proportion of alleles across both data sets (cases and controls), calculated with Equation 6.2. Across both datasets there were a total of 1676 alleles.

Case reference allele = $\sum all\ reference\ alleles \times (668 \div 1676)$

Case alternate allele = $\sum all\ alternate\ alleles \times (668 \div 1676)$

Control reference allele = $\sum all\ reference\ alleles \times (1006 \div 1676)$

Control alternate allele = $\sum all\ alternate\ alleles \times (1006 \div 1676)$

Equation 6.2: Equation used to calculate the expected values for the Chi-squared test

Chi-squared values were calculated for each allele in both the control and case data (so that there were four chi-squared values per chromosomal position variant), Table 6.3 and

Equation 6.3. These four different values were added to give the total chi-squared value for that variant (n). R (The R Foundation, Vienna, Austria) was used to calculate the chi-squared and p-value for each variant [248].

Chromosomal position	Allele A	Allele T	Total
Cases	a	b	a+b
Controls	c	d	c+d
Total	**a+c**	**b+d**	**n**

Table 6.3: Contingency table showing the two possible alleles at a single chromosomal position, allele A and allele T.

$$\chi^2 = \sum \frac{(\boldsymbol{observed_i} - \boldsymbol{expected_i})^2}{\boldsymbol{expected_i}}$$

Equation 6.3: Formula for Chi-squared calculated separately for each allele at each variant in both cases and controls. *i* is either the case or control allele values.

The chi-squared distribution can overestimate for small (2x2 contingency table) data sets and values <5. There are 2 data sets (case major allele and case minor allele compared to control reference allele and control alternative allele) so the Yates correction was applied to the values. The Yates correction subtracts 0.5 from the absolute calculation (observed minus expected) [249].

The results of the chi-squared test were sorted on ascending order of p-value and the p-values for all variants were used for the association plots.

6.1.2. Computational analysis: assessing functionality of variants

6.1.2.1. Linkage disequilibrium

LocusZoom (version 1.3, GPLv3) was used to plot the variant p-values against chromosomal position with data on control linkage disequilibrium (LD) [250]. For this a reference SNV is compared to all the other SNVs in the region and the colouring of the plot is representative of the square of the correlation coefficient (r^2) (calculated in PLINK which uses an expectation-maximisation algorithm [251]). LD was calculated

with the control VCF file from the phased v3 1000 genomes project in PLINK v1.9 [252].

```
plink --vcf 1kgpv3_chr2 --r2 --out 1kgpv3
```

The input plot data is the variant chromosomal position and the p-value with the matching European control LD data, example shown in Table 6.4.

Variant position	Negative logarithmic p-value	r^2
160856173	227	
160833188	2	0.99
160839341	4	0.99

Table 6.4: Input plot data for LocusZoom plots, variant position, p-value and correlation co-efficient.

Plots were created for all variants irrespective of their function. A second plot was created for exonic (synonymous, non-synonymous, nonsense and splice sites) and intronic regulatory region (TFBS, splice sites and 5' and 3' untranslated regions) variants. A final plot was created for the exons and splice site variants to appreciate LD of the lead variant with these components alone.

To obtain the functional annotations, custom annotations were generated utilising the UCSC variant annotation integrator tool [253]. The selected annotated variants include all protein coding gene transcript effect predictions and regulatory annotations. Specifically, the protein coding gene transcript effect predictions can differentiate between intergenic variants, upstream and downstream variants, 5' and 3' untranslated region (UTR) variants, synonymous and non-synonymous variants, in-frame indels, frameshift variants, variants in the stop or start codons, nonsense variants, splice region variants (1-3 bp of exon or 5-8 bp of intron) or intronic

variants. TFBS data was extracted via UCSC from the transcription factor ChIP-seq table sourced ENCODE data [254].

6.1.2.2. Isoform identification

RefSeq (NCBI Reference Sequence database, Bethesda, USA) identifies three different isoforms for *PLA2R1* messenger ribonucleic acid (mRNA) transcripts; variant 1 (NM_007366), 2 (NM_001007267) and 3 (NM_001195641) [255]. Variant 1 is the longest isoform at 5633 bp and the longest transcript which was used in all subsequent analyses. Transcript variant 2 has a splice variant and is for the secretory isoform of PLA2R1 (whereas I am interested in the transmembrane isoform) with 5175 bp. Transcript variant 3 has 5627 bp via an alternate in-frame splice site creating a shorter protein, Figure 6.3.

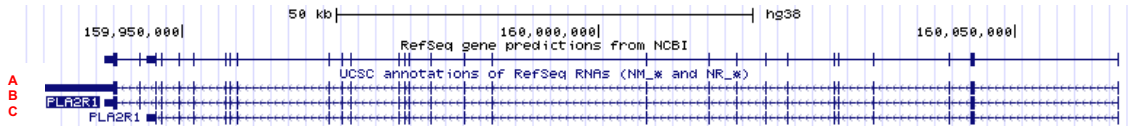


Figure 6.3: Graphical representation of the different PLA2R1 isoforms A. isoform 1, B. isoform 3, C. isoform 2.

6.1.2.3. Coding variants

Detailed functional annotation was undertaken of the 2 coding variants with a p-value $<5 \times 10^{-8}$. Multiple online resources were utilised to maximise successful identification and function of the variant; UCSC genome browser reference sequence hg19, NCBI (National centre for biotechnology information) dbSNP database, InterVar, ClinVar, OMIM (online Mendelian inheritance in man), ExAC (exome aggregation consortium), HSF (Human splicing finder), and variant effect predictor on ENSEMBL [237, 238, 256-261].

6.1.2.4. Regulatory region variants

Functional effect of significant (p-value <5 x10^{-8}) regulatory region variants within a TFBS was examined utilising a multitude of databases including those above. The name and motif of the TFBS data was obtained from the ENCODE transcription factor ChIP-seq data on the UCSC website [262]. TRANSFAC public domain and the associated programs such as Patch 1.0, P-Match 1.0, MatrixCatch 2.7, Alibaba 2.1, SignalScan, Match, were used to predict alternative TFBS DNA motifs [263, 264]. HSF 3.0 was used to predict the variant influencing a splicing site [261].

6.1.3. *In vitro* analysis: electrophoretic mobility shift assay

To determine the functionality of the lead variant on TFBS an *in vitro* method - electrophoretic mobility shift assay (EMSA) - was undertaken [243]. This required different steps, detailed below.

6.1.3.1. Transcription and translation of *CEBPB*

Dry *CEBPB* plasmid DNA (Origene, USA) was purchased as 10μg of dry powder. This was reconstituted as per manufacturer's instructions with 100μL of 18.2 mΩ water. The concentration of the *CEBPB* DNA was checked using a nanodrop machine and was 31.4ng/μL.

The TNT Quick coupled transcription/translation system (Promega, USA) was used to produce the recombinant CCAAT/enhancer binding protein beta (CEBPB) protein

[265]. The TNT Quick coupled transcription/translation system is a quick single tube reaction to produce cell free protein. It combines RNA polymerase, nucleotides, salts, amino acids and a recombinant RNAsin-ribonuclease-inhibitor. The expressed protein can be used directly after expression and no additional protein purification is required. An empty DNA vector was also created for a control using the same method and were processed together. The TNT single master mix was incubated with 1µl methionine and 100ng of the plasmid DNA. The product size was confirmed with 1-dimensional electrophoresis using a 12% Bistris pre-made gel (ThermoFisher Scientific, USA) and Western blot at 200 volts (V) electrophoresis. Visualisation was done by chemiluminescence with antibodies from mice horseradish peroxidase. A comparison was made to a genomic ladder with known molecular weights. The predicted weight of the recombinant CEPBP product was 39.71 kDa.

6.1.3.2. Synthetic DNA oligonucleotides

Synthetic oligonucleotides of 100bp with and without the lead variant were designed and purchased (Integrated DNA Technologies, USA), Figure 6.4. These were reconstituted with fresh water to achieve a 100µM concentration.

a. 5’ -
AGCCACCACGCCCGGACTACGTAATTTTTAAATGTCCCTTGTCATACAAATGCCTTGT
AAAAGTTCATTAAATGAATGGCTTTAATTATGCAATAGGGTT – 3’
b. 5’ –
AGCCACCACGCCCGGCCTACGTAATTTTTAAATGTCCCTTGTCATACAAATGCCTTGT
AAAAGTTCATTAAATGAATGGCTTTAATTATGCAATAGGGTT – 3’

Figure 6.4: DNA sequence of designed synthetic oligonucleotides measuring 100bp for lead variant in *PLA2R1*. a. Control oligonucleotide sequence b. Oligonucleotide sequence of region of interest in *PLA2R1* with the lead variant highlighted in yellow.

6.1.3.3. EMSA

To determine functionality and effect of the variant on CEBPB transcription factor binding an EMSA was done. The EMSA product was a combination of CEBPB and synthetic DNA oligonucleotides (with and without the variant). The shift assay was performed with the Lightshift chemiluminescent EMSA kit (ThermoFisher Scientific, USA), Table 6.5.

A 4% acrylamide non-denaturing gel was made using 892.15mL water, 5mL acrylamide, 2.5mL 10X Tris-borate-EDTA buffer. This was degassed for 15 minutes then 150μL 10% ammonium persulfate and 50μL of tetramethylethylenediamine. This was left to polymerise and then poured in to 1.5mm cassettes with 10mL each.

The EMSA products were run through the 4% acrylamide gel at 100V. Then transferred via immunoblotting to a polyvinylidene fluoride (PVDF) transfer membrane, fixated with UV transillumination and then visualised with chemiluminescence.

Order and Names of Chemicals	**Volume or Concentration**
Milli-Q 18.2Ω water	12 µL
10x binding buffer	2 µL
50% glycerol	1 µL
100Mm $MgCl_2$	1 µL
1µg/Ul Poly (dI.Dc)	1 µL
1 % NP-40	1 µL
Recombinant CEBPB	2 pmol
Wait 15 minutes	
Biotin labelled target DNA (working stock = 1 pmol/µL)	2 µL
Wait 20 minutes	
Loading buffer	5 µL

Table 6.5: EMSA protocol in the sequence of chemical addition and volumes and concentration used

6.1.4. In vitro analysis: replication

6.1.4.1. KASP genotyping

KASP genotyping was outsourced to LGC Genomics. The source reference genome sequence was provided to LGC with 50 bp up and downstream of the single lead variant of interest. The primer design and genotyping was done by LGC and they provided the results.

6.1.4.2. Illumina SNP Microarray

Due the cost of the alternative methods for replication and the failure of KASP genotyping I investigated if a pre-existing microarray Illumina beadchip already

contained the two lead variants of interest. The manifest of the following beadchips was reviewed: Omni 2.5, CytoSNP-850k, Global Screening Array and the Multi-ethnic global array. None included the variants of interest. For a custom beadchip the minimum number of samples is 1152, thereby being too expensive [266].

6.1.4.3. Hybrid Capture sequencing

New England Biolabs' (NEB) new technology NEB Next Direct is an alternative and more efficient way of hybrid capture sequencing. This technique was sought for the top ten lead variants, but of these only 7 were covered with the primer design and the cost was £160 per DNA sample. More traditional methods of hybrid capture sequencing with long range PCR were investigated for the entirety of the *PLA2R1* gene with Nonacus (UK). At a cost of £120 per DNA sample only two of the ten lead variants were adequately covered. Primer design alterations did not improve the coverage further, therefore this was not pursued as an option.

6.1.4.4. Sanger sequencing

Due to failure and cost with methods above an alternative was sought with the possibility of Sanger sequencing for the top two variants. This was cheaper at £2.60 per DNA sample.

6.1.4.4.1. Primer design

PCR primers were designed with Primer-BLAST (NCBI) [267, 268]. A combination of 2 sets (forward and reverse) of primers were designed first for each variant (number 1,2,8 and 9). When these did not work, after re-analysis further PCR primers were developed and ordered. PCR primers were ordered and bought from Integrated DNA

technologies (USA). The full set of primers ordered are summarised in Table 6.6. The position of the primers was visualised with SnapGene software v4.0.0 [269], see Figure 6.5 for the lead variant and Figure 6.6 for the cluster of second variants.

Table 6.6: Table of primer combinations designed and trialled for variants 1 and cluster of variants in 2

No	SNV	Forward strand sequence 5' to 3'	Reverse strand sequence 5' to 3'
1	1	CCACTAGGAAGGTAGTTTGGCTA	CCCAGATTATAGCAACATGAAGGC
2	1	TGCTGGTTATTTCTCATAGCACC	TTTTTTTGAGACGGAGTCTCGCT
3	1	CTGCTCAATCTTTGAGATGT	ATCCGCCCGCCTCGGCCTC
4	1	CATTTGTATGACAAGGGACA	GATGGTCTCGATCTCCTGAC
5	1	TTCTATAACCCTATTGCATA	ACATGAAGGCATGCACATTAC
6	1	CACATGTATTCAGTCTACTGCTCAA	TTTGCTACCACTCCCAGA
7	1	AATCTTTGAGATGTCTCTCTCTGTG	TCACCACTCCCAGATTATAGCAA
8	2	ACATGCTAAGGCAGTGTCTCT	AACATACCCATGCCTGTCAATAAGA
9	2	AAACAATACCTCTAAATGTCT	GGCATGAGAATCTCTTGAATC
10	2	ACATGCTAAGGCAGTGTCTCT	AACATACCCATGCCTGTCAATAAGA
11	2	TGGGTCTAGCTCTGCTGAAA	AATCCAGGAGGCAGAGGTT
12	2	AAGGAGAAACAATACCTCTA	CAGTAAGCTGAGATTGCAC
13	2	AGGGGGTTGTTGGGATTTGG	TTACGGCAGCATTGTCCTGA

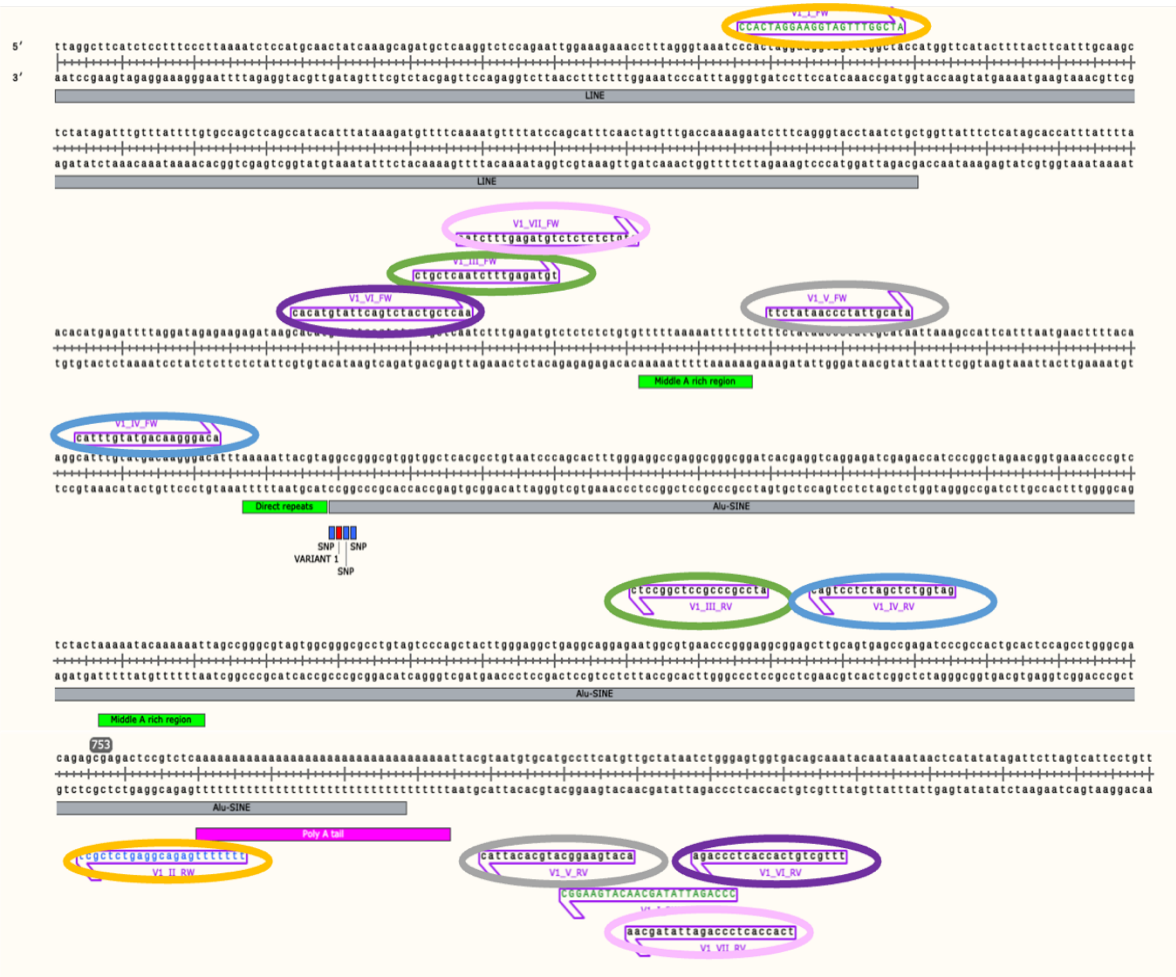


Figure 6.5: Annotated PCR primer location in comparison to the lead variant in *PLA2R*. Matching colours are the matching forward and reverse primer combination. Regions that pose specific PCR challenges are shown with the solid bars underneath the sequence.
Key: Green = regions of repeats, grey = interspersed nuclear elements (either long or short), pink = poly-A tail.
Alu-SINE = Alu-short interspersed nuclear elements, LINE = long interspersed nuclear elements

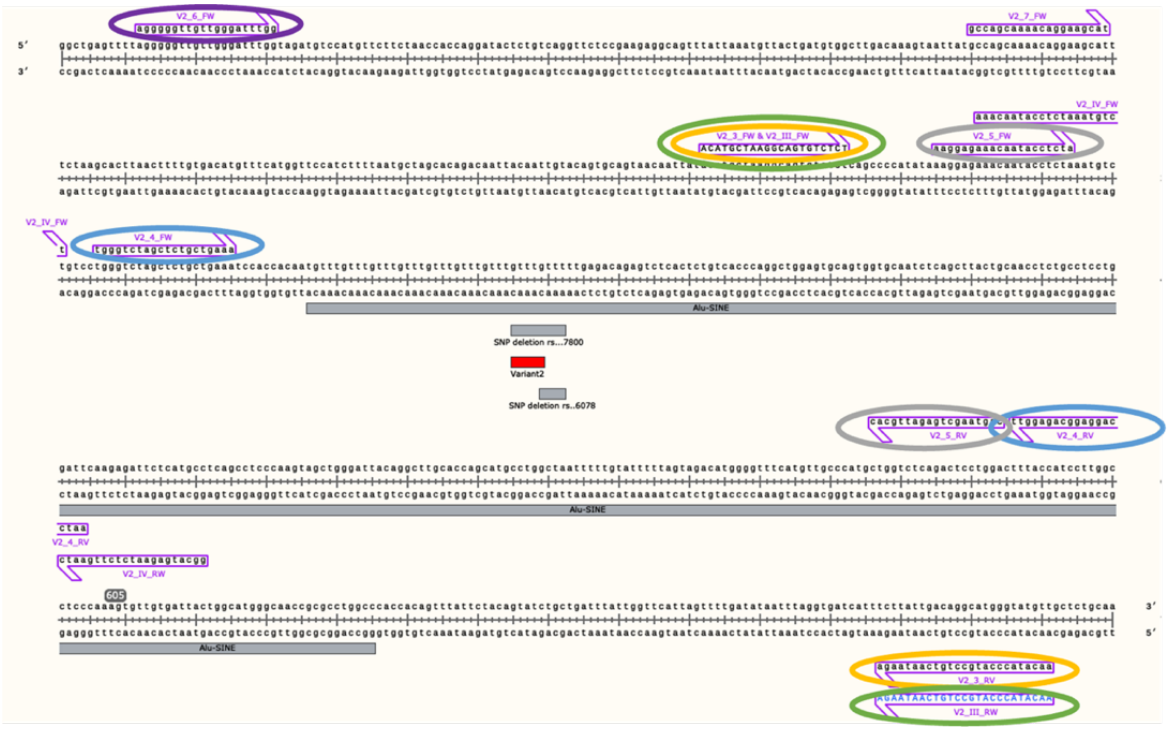


Figure 6.6: PCR primer location in comparison to the cluster of second lead variants in PLA2R. Matching colours are the matching forward and reverse primer combination.
Key: Green = regions of repeats, grey = interspersed nuclear elements (either long or short), pink = poly-A tail.
Alu-SINE = Alu-short interspersed nuclear elements, LINE = long interspersed nuclear elements

6.1.4.4.2. PCR

Standard PCR protocols were followed with a combination of whole genomic control DNA (250-500ng), 10µM of both each forward and reverse primer, and 12.5µL of ReadyMix Taq PCR reaction mix with $MgCl_2$ (P460, Sigma Aldrich, USA) to a total volume of 25µL. This was then amplified in the thermal cycler (DNA Engine Peltier thermal cycler, PTC-200, Bio-Rad, USA), Figure 6.7.

PCR was performed with the different matching primer forward and reverse combinations with two different healthy control whole genomic DNA samples. Different primers and combinations were tested and the recommended annealing

temperatures were used for each primer set, this was either 55º or 52º. Because the trial for the first 2 sets of primers was unsuccessful, I also tried nested PCR, whereby the outer region was first amplified and then using inner primers the same DNA was re-amplified.

Production of the correctly sized PCR products was examined with electrophoresis using a 1% agarose gel at 60V. The PCR product was visualised by UV transillumination using ethidium bromide. DNA band sizes were compared against a DNA ladder control (Quickload 100bp DNA ladder, New England Biolabs, UK).

Figure 6.7: A flow chart of the thermal cycler settings and steps

6.1.4.4.3. Purification of DNA

If the PCR DNA product was the expected product size it was extracted from the agarose gel by cutting the band under UV transillumination. The nucleic acid product was extracted with the High Pure PCR Product Purification Kit (Roche, Sigma Aldrich, USA). The standard protocol, Table 6.7, was followed to extract the PCR DNA [270]. The DNA concentration of the PCR product was measured using the Nanodrop ND-1000 (Labtech, UK).

Step	Instruction
1	Add 500µL binding buffer → Thermoshock 500 RPM (revolutions per minute) for 10 minutes
2	Add 300µL Isopropanol and then vortex
3	Transfer 600µL onto the filtered Eppendorf tube (from the kit)
4	Centrifuge 12000RCF (relative centrifugal force) at 21º for 1 minute
4	Transfer remaining effluent from (2) onto filtered Eppendorf tube → repeat (4)
5	Add 500µL wash buffer → Centrifuge 12000RCF at 21º for 1 minute
6	Add 200µL wash buffer → Centrifuge 12000RCF at 21º for 2 minutes
7	Transfer effluent to clean collection Eppendorf
8	Add 50µL elution buffer → Wait for 5 minutes
9	Centrifuge 12000RCF at 21º for 2 minutes → Pure PCR product made

Table 6.7: High Pure PCR product purification protocol, buffers are provided with the kit.

6.1.4.4.4. Sequencing

Sequencing was outsourced to GATC Biotech (Germany) by purchasing LIGHTRUN tube barcodes. Once the DNA was extracted 5µL of the PCR product was mixed with 5µL of the forward primer in one Eppendorf and 5µL of PCR product with 5µL of the reverse primer in the other, as per the GATC Biotech protocol [271]. The premixed samples of control DNA and primers was sent for Sanger sequencing [271]. Results were sent back for analysis electronically. Electropherograms were visualised using the ab1 file and the 4Peaks program [272].

There was no signal for either variant with either of the primers in the electropherograms, see Figure 7.13. Thus, the nested primers were tried and again this failed, Figure 7.14. At the third attempt I used a combination of nested PCR primers and exclusively used the pre-amplified control DNA so that the overall concentration of DNA would be higher and reduce non-specific binding elsewhere within the genome.

6.2. Genome wide association study

6.2.1. Computational tools

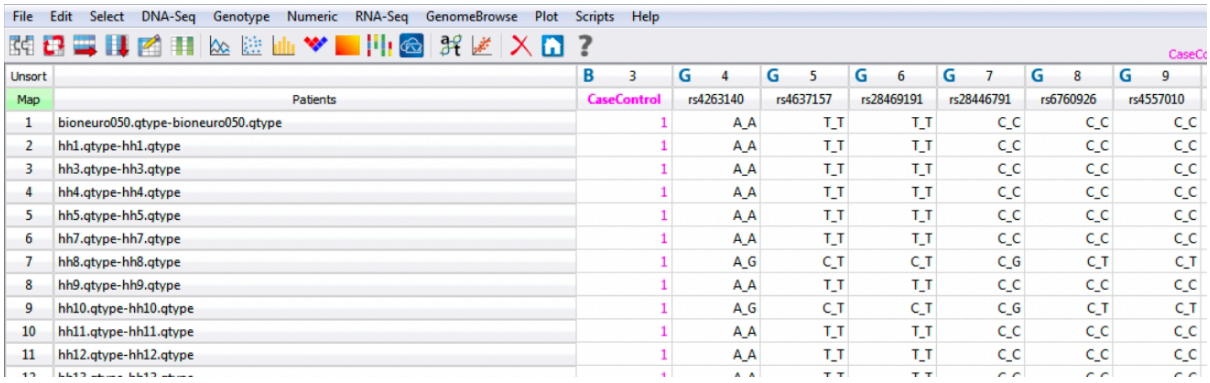
6.2.1.1. PLINK

PLINK is a free, open-source whole genome association analysis toolset. It is the most used toolset for a range of basic and large-scale dataset analyses. Most analyses were undertaken using PLINK v2.0 alpha and for some functions that were not available in v2.0, PLINK v1.90 beta was used [252, 273-276]. The advantage of v2.0 compared to v1.9 despite it being in developmental stages is that the processing speed is much faster, and tools are available for larger datasets that require a lot of computational processing power. PLINK is a command line program with no graphical user interface. Input file formats are different depending on the version used but are in binary formats to improve processing speeds. A single file with binary data for the genotype information is made, called a *.bed or *.pgen in v1.9 and v2.0 respectively. The binary file *.bim or *.pvar contains identifier information about all of the genotyped SNVs. The sample information is contained within the *.fam or *.psam file and can also be used to add phenotype information to the dataset [277].

6.2.1.2. SNP & Variation Suite

Golden Helix SNP & Variation Suite version 8.8.1 (SVS) is a specialist subscription paid for software package designed to analyse genomic data [278]. It has a user-friendly interface with drop down menus and selections and does not require any coding and the genotyping data can be viewed easily. SVS uses a *.dsf format for

files and can import and export to a variety of different formats improving its functionality. Once imported, genotyping data is visible on screen with columns for the SNV markers and a different row for each individual, Figure 6.8 [278].



Unsort		B 3	G 4	G 5	G 6	G 7	G 8	G 9
Map	Patients	CaseControl	rs4263140	rs4637157	rs28469191	rs28446791	rs6760926	rs4557010
1	bioneuro050.qtype-bioneuro050.qtype	1	A_A	T_T	T_T	C_C	C_C	C_C
2	hh1.qtype-hh1.qtype	1	A_A	T_T	T_T	C_C	C_C	C_C
3	hh3.qtype-hh3.qtype	1	A_A	T_T	T_T	C_C	C_C	C_C
4	hh4.qtype-hh4.qtype	1	A_A	T_T	T_T	C_C	C_C	C_C
5	hh5.qtype-hh5.qtype	1	A_A	T_T	T_T	C_C	C_C	C_C
6	hh7.qtype-hh7.qtype	1	A_A	T_T	T_T	C_C	C_C	C_C
7	hh8.qtype-hh8.qtype	1	A_G	C_T	C_T	C_G	C_T	C_T
8	hh9.qtype-hh9.qtype	1	A_A	T_T	T_T	C_C	C_C	C_C
9	hh10.qtype-hh10.qtype	1	A_G	C_T	C_T	C_G	C_T	C_T
10	hh11.qtype-hh11.qtype	1	A_A	T_T	T_T	C_C	C_C	C_C
11	hh12.qtype-hh12.qtype	1	A_A	T_T	T_T	C_C	C_C	C_C

Figure 6.8: Display of an example of the format and layout of genotyping data in SNP & Variation Suite

6.2.1.3. VCFtools and BCFtools

VCFtools is a package containing a set of utilities that manipulate genotyping files in the *.vcf format and the binary counterpart *.bcf is manipulated using BCFtools. It is a command line tool and facilitates data filtering and reading and writing of a VCF file [279, 280]. The VCF file is a standard format for storing genetic variation data and is a tab delimited text file that contains genetic information [280, 281]. The headers include meta-information such as the data source, the date and the format of the genotypic data. This information is ignored in the analysis of the VCF file as it is prefixed by a hashtag '#' [281]. The genotyped data itself is organised by the chromosomal position in rows and each separate individual in the columns [281]. At each variant for each individual the genotyped data is contained with a 1 or 0 format to depict if the variant is present or not [281]. This way the data is represented in a matrix depicting the variants each individual has, Figure 6.9.

```
##fileformat=VCFv4.2
##fileDate=20191004
##source=PLINKv1.90
##INFO=<ID=PR,Number=0,Type=Flag,Description="Provisional reference allele, may not be based on real reference genome">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
#CHROM  POS ID  REF ALT QUAL    FILTER  INFO    FORMAT  bioneuro050.gtype   hh1.gtype   hh3.gtype   hh5.gtype   hh7.gtype
1   798959  rs11240777  G   A   .   .   .   GT  1/1 0/1 0/1 0/0 0/1 0/0 0/1 0/1 0/1 0/1 0/0 0/0 0/0 0/0 1/1 0/0 0/0 0/1 0/0 0/0 0/0
1   838555  rs4970383   G   T   .   .   .   GT  0/0 0/1 0/0 0/0 0/0 0/0 0/0 0/1 0/1 0/0 0/0 0/0 0/0 0/0 0/1 0/0 0/1 0/1 0/1 0/0
```

Figure 6.9: An example of the first few lines of a variant call format file, showing the format of the genotyping data

6.2.1.4. Beagle and SNP2HLA

Beagle 5.1 is used for whole genome phasing and imputation of missing genotypes [215]. Ungenotyped SNVs in the study sample are imputed if they are present in the reference panel. Beagle clusters haplotypes at each locus and the number of clusters adapt to the amount of information available and increase globally with sample size and locally with linkage disequilibrium [282]. Beagle is ideal for the study as it is suitable for larger sample sizes >1000 individuals [282]. It can improve computation processing time by clustering haplotypes and increasing processing with sample size. Version 5.1 provides faster algorithms and uses a new reference format called bref3. Beagle uses Java version 1.8 as a minimum (Java can be downloaded from www.java.com) [282].

SNP2HLA was developed by the Broad Institute and is publicly available. SNP2HLA uses Beagle to impute HLA alleles and the amino acid polymorphisms using reference haplotype data [225]. SNP2HLA requires access to PLINK and an older version of Beagle (version 3.0.4). HLA imputation is different to whole genome imputation because it exclusively focuses on the HLA region on chromosome 6 which is highly polymorphic and has complex linkage disequilibrium. As a result the pattern recognition algorithms are adjusted [223]. HLA imputation predicts the untyped HLA alleles based on the reference panel. Compared to other tools

available SNP2HLA can impute a greater number of alleles and the call rate is higher [226]. An appropriate reference panel is required to facilitate reliable imputation quality, see 6.2.7.1 for more information.

6.2.1.5. Remedy

Our in-house software Remedy [283] developed by Dr Cheshire converts Illumina output report files in to useable VCF files [283]. The input data is the *.txt Illumina final report or matrix format file. Remedy scans the microarray chip manifest file and matches each SNVs refSNP identifier (rsID) to the reference database dbSNP version 150. Using dbSNP as the reference it filters out non-matching SNVs, multiallelic variants, structural variants and erroneous SNVs. The output from this is the number of valid SNVs that are useable for further analysis. Remedy scans the genotyping data and detects the encoding system across the whole dataset using the human genome build 37. Remedy can then convert to a desired encoding system output of your choice and outputs the file in to a workable VCF file.

6.2.1.5.1. Encoding schemes

The encoding system states the strand of DNA on which the SNV lies. There are different encoding systems that have been developed.

1. Forward and reverse: This scheme is based on the forward (5’ to 3’) strand in the reference genome. This uses the nucleotides ACGT letters to encode the alleles. It is the most frequently used encoding scheme because it is based on the reference genome. In most instances this is the same as the encoding scheme used by dbSNP. It is important to specify the genome build used as the alleles can vary between different versions [284, 285].

2. Top and bottom: This scheme was developed by Illumina as a method to determine unambiguously the specification of the strand based on the surrounding DNA sequence without reference to any database or reference genome. The SNV can be identified and placed in reference to the surrounding DNA sequence and is encoded A or B. The strand designation of the SNV column is available in the manifest. This can be particularly useful when no reference is available [284, 285].

3. Plus and minus: The plus and minus scheme was predominantly used by the HapMap project (www.hapmap.org). The plus strand corresponds to the genomic sequence as for the genomic reference and so again is dependent on a specific reference panel and a particular version of it [284, 285].

It is important that when comparing alleles in association studies that the encoding schemes match particularly in instances where the different datasets have been genotyped and the data files generated differently. Therefore, Remedy is vital so that the mixed datasets used in this study can be converted to a common and directly comparable encoding scheme.

6.2.2. Case sample

6.2.2.1. Consent and inclusion criteria

The case dataset comprised of individuals with biopsy-proven AMN. Secondary causes had been excluded by their consultant nephrologist. Ancestry data was not readily available. Written informed consent was obtained by each collaborator at each site in accordance with local guidelines. A summary of centres and sample

numbers are shown in Table 7.3. Individuals from the North East & Central London cohort include AMN cases from the Royal Free NHS Trust & Royal London Hospital. These individuals were recruited by myself via the tertiary membranous clinic at these sites.

6.2.2.2. DNA extraction

Whole blood samples were obtained from consenting AMN patients across multiple collaborating European centres. Each individual was assigned a unique anonymised identification code. Anonymised whole blood samples were sent to UCL Genomics (UCL Great Ormond Street, Institute of Child Health, London) for DNA extraction. Extracted DNA samples were returned, then I plated each individual DNA sample on to 96 well plates with 200ng of DNA in each well for genotyping.

6.2.2.3. SNV microarray sequencing

Genotyping of cases was undertaken at UCL Genomics (UCL Great Ormond Street, Institute of Child Health, London) on the Illumina Infinium Multi-Ethnic Global BeadChip with 1,779,818 markers. The microarray chip was designed specifically for a multi-ethnic cohort [286]. Sample processing for Illumina was done in accordance with the Infinium HD Ultra Assay protocol (Illumina Inc., San Diego, USA) [287, 288]. Whole genomic sample DNA is transferred to a deep well plate, with the addition of reagents this is then amplified for 20-24 hours. The sample DNA is fragmented, precipitated and resuspended. As described in 5.4.1.2, DNA is hybridised to the microarray chip and this is left to incubate for a further 16-24 hours to ensure adequate hybridisation. Non-hybridised DNA is washed off and the microarray chip is prepared for staining. A single nucleotide complementary extension of the sample

DNA is made on the microarray chip. DNA fragments of non-complementary binding are washed away. The nucleotide is fluorescently labelled and the microarray chip is scanned and read using an iScan scanner (Illumina Inc, San Diego, USA). This is done through multiplexing to undertake high throughput sequencing, see 5.4.1.3. At UCL Genomics the advantage is that this whole process is automated using the liquid handling robot (Freedom Evo, Tecan Ltd, Switzerland) so even though each microarray chip can only process 96 samples and the time to sequencing is 3 days the automation greatly facilitates sequencing of multiple samples together. Data from the iScan scanner is generated in raw intensity files (IDAT) format and these files are provided, see 5.4.1.2.1 for more information.

6.2.3. Case data preparation

6.2.3.1. Raw data to report file

IDAT files were provided for each 96 well plate as the output from microarray sequencing. These had to be converted to useable genotyping data format for further downstream analysis. Illumina provide open access to GenomeStudio software v2011.1 which allows conversion of IDAT files to a useable format [289]. GenomeStudio v2011.1 runs on a Windows interface. GenomeStudio needs the *.csv manifest file, which is available via: http://emea.support.illumina.com/array/array_kits/infinium-multi-ethnic-global-8-kit/downloads.html?langsel=/gb/ (Infinium Multi-Ethnic Global-8 v1.0 Manifest File Build 37 - CSV Format). Genome build 37 was chosen as Remedy works in genome build 37, see 6.2.1.5. The cluster file *.egt was also downloaded from the same website (Infinium Multi-Ethnic Global-8 v1.0 Cluster File) and the location input to GenomeStudio. The first filter to be applied in GenomeStudio is the GenCall

threshold. The GenCall score is a metric ranging from 0 to 1 for each genotype, the further a sample is from it's associated cluster of data the lower the score. Illumina recommend this is set at 0.15 to exclude genotypes that are too far from the cluster centroid and are therefore unreliable [290]. GenomeStudio combines the IDAT files with the bead pool manifest file and creates a file called a genotype call file. This needs to be done separately for each 96-well plate of DNA. There are different encoding schemes that can be chosen and different formats for the genotyping data, see 6.2.1.5.1. The report wizard tool was selected from analysis and reports. AMN samples were selected from the 96-well plate and the matrix report format was chosen to generate the report in a *.txt file format with tab delimiting spaces. This was repeated for all 96 well plates separately.

6.2.3.2. Report *.txt file to PLINK file

Our in-house software Remedy [283] converts the output report files in to useable VCF files [283], see 6.2.1.5 for further information. The purpose of Remedy is to filter and ensure correct strand encoding for the genotyped data. For the purposes of this analysis I wanted the data in the forward encoding system. The forward / reverse encoding system is the same as dbSNP and most routinely used. This encoding system states the strand of DNA on which the SNV lies – either the forward or reverse strand. Remedy excludes; loci with multiple mappings; pseudo-autosomal regions; and no probe mappings. It then scans the microarray chip manifest file and attempts to match each SNVs rsID to the reference database dbSNP version 150.

Each 96 well plate *.txt report file is processed separately thorough Remedy. Therefore, there are multiple different VCF files produced from Remedy and these

need to be merged to make a single case dataset. Initially the files were converted to PLINK format bfiles. Merging the datasets did not work because multi-allelic variants remained. These were filtered out using VCFtools and a merge of the different sub-datasets was attempted using VCFtools. Ambiguous SNVs that remained within the dataset were excluded and filtered out using SVS. The VCF file was recoded using PLINK v2.0 [252, 273].

6.2.4. Control data preparation

6.2.4.1. Downloading publicly available datasets

Three different sources of publicly available control datasets were used to obtain an overall large enough control cohort; the Oxford cohort, the Illumina cohort, and the Wellcome Trust Case Control Consortium Controls (WTCC), the data is summarised in Table 6.8.

The Oxford cohort is accessible through the European Genome-Phenome Archive (EGAD00010000144 and EGAD00010000520) [291, 292]. The dataset has 432 self-reported European healthy volunteers genotyping data. Genotyping was done on a HumanOmniExpress-12 v1_J microarray chip with 730,525 SNV markers in 144 individuals. In the remaining 288 genotyping was done on HumanOmniExpress-12 v1_A with 733,302 SNV markers. Overlapping SNV markers on both microarray chips totalled 730,397. The Illumina ethnicity controls are accessible through Illumina [293]. The dataset contains 270 individuals across 4 different ethnicities, the 90 Central European individuals were extracted. Genotyping is performed on a HumanOmniExpress-12 v1_C microarray chip with 731,442 SNV markers.

The Wellcome Trust Case Control Consortium controls 2 (WTCC) is a combined dataset of the 1958 British Birth Cohort and the UK National Blood Donor control group. Data is available through the WTCC website [294]. The 1958 British Birth Cohort is a control set of 2,867 self-reported White individuals born in England, Wales and Scotland in 1958 and subsequently followed up for a maximum of 42 years [295]. The remainder of this dataset comes from the UK Blood Service and is from 2,737 healthy blood donors. Genotyping of the WTCC dataset was performed on an Illumina 1.2M Duo Custom BeadChip with 1,106,184 SNV markers.

Dataset	Sample number	Ethnicity	Microarray Chip	SNV marker count
Illumina	90	EUR	OmniExpress	731,442
Oxford	432	EUR	OmniExpress	730,397
WTCC	5604	EUR	Duo Custom	1,106,184

Table 6.8: Summary of control datasets.
Key: EUR =European, WTCC =Wellcome Trust Case Control Consortium controls

6.2.4.2. Processing to workable PLINK file

The control datasets were downloaded in a compressed format and were extracted. The encoding schemes of the Illumina and Oxford datasets were in the desired forward / reverse encoding system but the WTCC dataset was not. For this reason all three control datasets were processed through our in-house software Remedy [283]. This was also done for the Illumina and Oxford datasets as Remedy has useful features of filtering other SNVs (including multi-allelic and structural variants) in addition to re-encoding. The outputted VCF file was recoded using PLINK v2.0 to PLINK workable files [252, 273].

6.2.5. Quality control

Quality control (QC) of the genotyping data is essential as genotyping is not perfect and so errors are bound to present. A multitude of errors can occur at each stage of the process such as poor quality DNA, contamination, poor hybridisation to microarray probes, or poorly performing microarray probes. Small artefactual errors can generate false positives and false negatives when a large sample size is being investigated. In this GWAS whereby the cases and controls are genotyped on different microarray platforms this is a source for another considerable error. The cases and controls must have strict QC measures so that they can be as closely matched to one another and so that any associations found are true.

The QC methods can be considered in three distinct sections, QC per individual, per SNV marker and population stratification. Methods are adapted from Anderson *et al.* [296]. QC was done separately for cases and controls because of the different genotyping microarray chips and across different centres. Once QC was completed then the combined case-control dataset was re-processed for further QC.

6.2.5.1. Filtering genotyping data

6.2.5.1.1. Per individual filtering

Filtering for each individual has three main components; identifying individuals with missing genotype data, discordant gender information and related individuals. The test for discordant gender information was not conducted because only autosomal chromosomes were analysed, see 5.4.2.1.

6.2.5.1.1.1. Call rate

The call rate helps exclude individuals that have a high rate of missing genotypes. It is the fraction of genotyped SNVs over the total number of SNVs in the dataset. This can be due to variation in DNA quality and concentration. A stringent threshold of excluding any individual with a total genotyping call rate of <98% was used. The filtering was done in PLINK using the --mind command.

```
plink2 --bfile mncases --mind 0.02 --make-bed --out
mncasemind0.02
```

6.2.5.1.1.2. Heterozygosity rate

The heterozygosity rate is the proportion of heterozygous genotypes that each individual has. A high heterozygosity means a high degree of genetic variability and for an individual this is suggestive of low sample quality. A low heterozygosity rate suggests little genetic variability and can be seen in consanguinity. Heterozygosity needs to be performed on SNV markers that are not highly correlated, so regions of high linkage disequilibrium were excluded and further pruning of the SNV markers using PLINK and the --exclude and --range tools was done. Heterozygosity checks were done on a pruned dataset of 98,676 independent SNVs. On the pruned dataset the heterozygosity rates were calculated using the --het tool. A visual review of the spread of heterozygosity rates in a graph enabled the decision to exclude individuals ±3 standard deviations from the mean. A text file was created using R and this was used to extract the remaining individuals in PLINK.

```
plink --bfile mncasemind0.02 --exclude inversion.txt --range -
-indep-pairwise 50 5 0.2 --out mncasemind0.02indepsnv
```

```
plink --bfile mncasemind0.02 --extract
mncasemind0.02indepsnv.prune.in --het --out mncaseld


plink2 --bfile mncasemind --remove het_fail_ind.txt --make-bed
--out mncasemindhet
```

6.2.5.1.1.3. Identity by descent

A GWAS analysis assumes independence of individuals, if this is not the case and individuals are related then there will be an inflated rate of false positives. Therefore, these individuals need to be excluded from the analysis such that only a single member from each family remains. Identity by descent (IBD) is a measure of the percentage of shared alleles at each SNV marker between each two individuals within the dataset [297]. The IBD score is calculated for each pair of individuals and reflects how related these two individuals are. The IBD score is imperfect as the genotyped data is unphased and so haplotypes are estimated. The more related an individual is the higher their score, so theoretically two unrelated individuals would share no alleles and would have a score of 0. Identical twins on the other hand would have an IBD score of 1 as they are genetically identical. Third degree relatives would have a score of 0.125.

For the purposes of this study, individuals with an IBD score >0.185 were excluded. Because of genotyping error, linkage disequilibrium and population structure there is some variation around these theoretical values and so an adjusted score of 0.185 is routinely used [296]. This is halfway between second- and third-degree relatives. In the pruned and independent SNV marker dataset I undertook IBD testing using the --genome tool in PLINK v1.9 [275]. PLINK uses a method to detect extended chromosomal segmental IBD between pairs of individuals using a hidden Markov

model and then uses a method-of-moments approach to estimate the probability of sharing of the alleles between the two individuals [274]. A perl script to examine the IBD scores was used which identified individuals to be excluded and they were removed using the --remove tool in PLINK v2.0.

```
plink --bfile mncasemindhet --extract
mncasemind0.02indepsnv.prune.in --genome full --out
mncasesmindhet
```

6.2.5.1.2. Per SNV filtering

To further minimise errors QC is conducted on each SNV marker in the genotyped dataset. This was started in Remedy using dbSNP as the reference, and filtered out non-matching SNVs, multiallelic variants, structural variants and erroneous SNVs. Further checks are warranted as this is only compared to dbSNP and not the genotyping. Further filtering per SNV marker is exclusion of the X and Y chromosomes, allele count filtering, call rate, minor allele frequency and Hardy-Weinberg equilibrium.

6.2.5.1.2.1. Exclusion of sex chromosomes

A decision was made to exclude the sex chromosomes from analysis. This is due to the difficulty in analysing data from the X chromosome and the coverage of reliable SNV markers for the X chromosome on SNV microarray chips (only 1.2% of all SNV markers were on the X chromosome in the Illumina Infinium Multi-Ethnic Global BeadChip. Further, imputation is not straightforward for the X-chromosome as it requires knowledge of gender and phased X-chromosome data and neither of these were available in the control dataset [298]. It is common in GWAS analyses to exclude the X chromosome because of the unique nature of X-inactivation and also

as men have only one copy it provides specific analytical challenges [299]. The sex chromosomes were filtered out by Remedy.

6.2.5.1.2.2. Allele count

Allele count is the different number of alleles that are possible at a single SNV. In association studies due to the statistics SNVs are chosen that are bi-allelic and so contain only 1 alternative allele. Multiallelic SNVs were filtered out by Remedy, dbSNP version 150. PLINK is not yet capable of handling multiallelic SNV data and so genotyped data cannot be imported unless it is biallelic. An extra check was done using BCFtools and the -m2 and -M2 commands which did not identify any unidentified multi-allelic SNVs.

6.2.5.1.2.3. Call rate

The call rate for the SNV is the percentage of individuals in whom genotyping of the SNV is available. If certain SNVs have high missing rates across all individuals it suggests issues during genotyping with that SNV marker and so should be excluded. I used a stringent threshold of excluding any SNV with a genotyping call rate of <98%. The filtering was done in PLINK using the --geno command.

```
plink2 --bfile mncasesmindhetibd --geno 0.02 --make-bed --out
mncasesmindhetibdgeno
```

6.2.5.1.2.4. Minor allele frequency

SNVs with a low minor allele frequency (MAF) are excluded. A low MAF means rare alleles which can result in false positives and often the association signal seen is dependent on a few individuals. The threshold of the exclusion depends on each sample size but most commonly SNVs with a MAF >5% are used in a GWAS to

examine common variants. A histogram for MAF was inspected and a decision made to exclude any SNV with a MAF <5%. This was done in PLINK v2.0 using the --maf command.

```
plink2 --bfile mncasesmindhetibdgeno --maf 0.05 --make-bed --
out mncasesmindhetibdgenomaf
```

6.2.5.1.2.5. Hardy-Weinberg equilibrium

The Hardy-Weinberg equilibrium (HWE) is a principle that states 'genotype frequencies in a population remain constant between generations in the absence of disturbance by outside factors' [300]. For a single locus with two alleles there are three possible genotypes AA (homozygote common allele), Aa (heterozygote) and aa (homozygote alternate allele). HWE states that the proportion of these remains constant and calculable within a population provided that there is random mating, there are no natural selection pressures, the population is infinitely large, no mutations occur and no emigration or immigration occurs to prevent gene flow [301].

Beyond these reasons a possible cause of deviation from HWE is sequencing errors [301]. Heterozygote excess can be seen, particularly in genomic regions such as segmental duplications and simple tandem repeats [301, 302]. The allele frequencies can be calculated in a genotyped dataset and the allele frequencies can be assessed for deviation from HWE and excluded accordingly. Deviations from HWE may well exist in the case dataset because of selection for the disease and the selection bias for the ascertainment of their DNA. Removing SNV markers in the case dataset that deviate from HWE could result in the loss of truly associated SNVs [296, 303]. As a result, HWE was performed on the control dataset alone. A high

threshold cut-off for SNVs out of HWE with a p-value <0.00001 was used. The filtering was done in PLINK using the --hwe command.

```
plink2 --bfile ctrlmindhetibdgenomeaf --hwe 0.00001 --make-bed
--out ctrlmindhetibdgenomeafhwe
```

6.2.5.1.3. Combined case-control dataset

Once the QC was done for all the above steps the case and the control datasets were merged. Further QC was done per SNV and intersecting variants were ensured prior to population stratification. The dataset was pruned to facilitate faster population stratification analysis.

6.2.5.2. Population stratification

Population stratification is a major source of confounding in genotype distribution and hence in case-control studies [296, 304]. Any small differences in the individual ancestry will highlight ancestral informative SNVs and not true disease association SNVs because allele frequencies for SNVs are different depending on the founder population [296, 304]. Population stratification and removal of individuals with divergent ancestry is necessary prior to any GWAS analysis and even subtle differences can exist in individuals from the same continent [296, 304]. In some studies it is possible to match ancestry in the case and control selection. In this study AMN cases were from three Western European cities; Manchester, UK; London, UK; Hamburg, Germany. The ancestry of cases is unknown and so is likely to represent a mix of the local population which in cities like London and Manchester is high. Further, in the control dataset these individuals were self-reported White European, which again may not be genetically accurate as there may be individuals with

admixed ancestry. To determine the identity of cases and controls the methods of principal component analysis (PCA) and multidimensional scaling (MDS) were used.

6.2.5.2.1. Principal components analysis

PCA is the most common method for identifying and removing individuals of different ancestry [296]. PCA is a mathematical algorithm that reduces the complexity of the data while retaining the variation within the dataset [305]. It identifies directions or independent variables (termed principal components) in a linear way from any data matrix containing observations across multiple variables [305]. The first principal component accounts for the greatest variation, and the proportion of variance reduces as the principal components increase [305]. This means that the greatest variation can be identified by the first and second principal components which can be visualised as a scatter plot. In genotyped data the observations are the individuals and the variables are the SNVs. A comparison is made to data from individuals with known ancestry. For the PCA analysis controls two datasets were available; 1. An Illumina dataset comprising 270 individuals; 89 African, 91 European and 90 East Asian and 2. the 1000 Genome Project controls.

PCA was performed using SVS on the case-control dataset with the Illumina controls as the reference ancestry. SVS uses the EIGENSTRAT method which was developed by the BROAD institute [306]. The EIGENSTRAT method applies the principal components analysis to genotype data and infers continuous axes of genetic variation [306]. This reduces the variation to a small number of dimensions defined as the top eigenvectors of covariance (same as principal components). As the ancestry information is a geographic interpretation the spread of the data

highlights the range of geographic spread of the data. For example an axis from northwest to southwest would have values that range as positive for the northwest to near zero for central and then to negative for southeast [306]. Due to limitations in computational power with SVS the analysis was limited to a random selection of 20,000 SNV markers; from previous work in our laboratory this has been proven to provide the same results as using all the SNV markers [307]. Further, the first 2 principal components should account for the majority of variance so only the first 10 principal components were calculated.

The advantage of SVS is that it has an in-built feature whereby outliers can be excluded. The method of outlier exclusion uses standard deviations cut-off at pre-specified chosen values. I examined the differences between the different standard deviation cut-offs visually in a graphical format of the PCA and from these chose the standard deviation cut-off of 2.5 with the Illumina ancestry controls, see Figure 7.25, Figure 7.26, Figure 7.27 and Figure 7.28 .

As an additional check to ensure correct extraction of exclusively European individuals with good overlap to the target imputation reference panel using the 1000 Genomes Project PCA was repeated using PLINK and an additional method called multidimensional scaling in PLINK was undertaken.

6.2.5.2.2. Multidimensional scaling

The MDS method detects underlying dimensions that explain observed genetic distance. MDS in PLINK and uses pairwise identity by state distance between each individual. Two advantages of MDS over PCA are; 1. it does not require the data to

follow a multivariate normal distribution and 2. It does not require computation of a covariance matrix and can be applied to any distance or similarity [308]. Below are the steps detailing how MDS was performed.

6.2.5.2.2.1. 1000 Genomes Project reference ancestry dataset

Individuals from the 1000 Genomes Project dataset for the reference ancestry were used and these were downloaded using the wget function on GNU command line through our in-house server [309]. The data is available through 'ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/release/20100804/ALL.2of4intersection.20100804.genotypes.vcf.gz'. This file from the 1000 Genomes Project contains genetic data of 629 individuals from different ethnic backgrounds. The VCF was converted to PLINK workable binary files. Data filtering was done using the same tools in PLINK as described in 6.2.5.1. Individuals with a call rate <98%, any SNV with an MAF <5% and any SNV with a genotyping call rate <98% were excluded. SNVs present in the combined case-control dataset were extracted from the 1000 Genomes Project dataset using 'awk' to generate a text list of SNVs present within the case-control dataset and extract these from the 1000 Genomes dataset in PLINK using the --extract tool.

```
awk '{print$2}' mn/casectr > unimputedmn/casectrlsnv.txt

plink --bfile 1kg --extract unimputedmn/casectrlsnv.txt --
make-bed --out unimputedmn/casectrlsnv_1kg
```

I also extracted the 1000 Genomes SNVs from the case-control dataset so that the SNVs matched exactly, the same process was repeated using 'awk' and --extract but the other way around. The two datasets need to have the same human genome build and genetic positions of the SNVs. To ensure this was the case the 1000

Genomes Project was changed to match the case-control dataset. This was done by creating a text file was from the case-control dataset as a reference map which contained the SNV identifiers and the physical position of that SNV. Using the --update-map tool in PLINK this text file was applied to the 1000 Genomes dataset to update the SNVs in this dataset.

```
awk '{print$2,$4}' unimputedmn/casectrlsnv.map >
unimputedmn/buildcasectrl.txt
```

```
plink --bfile unimputedmn/casectrlsnv_1kg --update-map
unimputedmn/buildcasectrl.txt --make-bed --out
unimputedmn/casectrl_1kg_hg
```

The next check was to make sure that the reference allele was matching within the two datasets. A text file was created using the *.bim file using the SNV identifier and the A1 allele (the minor allele). This was then applied to the 1000 Genomes dataset in PLINK using --reference-allele tool. Impossible minor allele assignments are excluded from any further datasets.

```
awk '{print$2,$5}' unimputedmn/casectrol_1kg_hg.bim >
unimputedmn/1kg_ref-list.txt
```

```
plink --bfile unimputedmn/casectrl_1kgsnv --reference-allele
unimputedmn/1kg_ref-list.txt --make-bed --out
unimputedmn/casectrl_refsnv
```

Differences between the two datasets may still exist, this could be due a strand issue or non-matching SNVs. Firstly, non-matching SNVs can be checked for by creating a text file of both datasets using the SNV identifier and the minor allele and the major allele. The differences between the two datasets are examined using the UNIX

command uniq -u which outputs a text file of the differences between the two datasets.

```
awk '{print$1}' unimputedmn/all_differences.txt | sort -u > unimputedmn/flip_list.txt
```

If after this discrepant SNVs remained these were removed from both datasets using the text file with the discrepant SNVs and the --exclude tool in PLINK.

6.2.5.2.2.2. Multidimensional scaling for PCA

The first step for MDS is creating the pairwise clustering matrix based on IBD. This helps detects pairs of individuals and their similarities and differences based on the genotyping data. To prevent overestimation of homogenous data this needs to be done in a minimum of 100,000 SNVs. This function can be done in PLINK using the --genome tool. Standard metric MDS is then calculated by analysing the matrix of the genome wide IBD pairwise distances. The --mds-plot tool allows you to specify how many dimensions you want extracted for each individual so that they can be plotted on a scatter plot, I chose 10 (as for PCA).

```
plink --bfile unimputedmn/mn1kgp_MDS_merge2 --read-genome unimputedmn/mn1kgp_MDS_merge2.genome --cluster --mds-plot 10 --allow-no-sex --out unimputedmn/mn1kgp_MDS_merge2
```

Using this file output and adding known ancestry information to the controls a scatter plot of the MDS data can be plotted in R, see Figure 7.32. An MDS covariate file can be created after exclusion of other ancestries and the same MDS step repeated, the covariate file can be used in the association analysis at a later step.

6.2.5.2.3. Genomic inflation factor

The genomic inflation factor is a way to quantify and measure the amount of stratification within a population [310]. It is calculated with the results of an association test and compares the observed and expected distribution of alleles [310]. Genomic inflation factor is also called a lambda score. Ideally the lower the genomic inflation factor the less the stratification, a factor as close to 1.0 is preferable [311]. A large genomic inflation factor can cause false positive associations.

To calculate the genomic inflation factor, the --adjust tool is used. This provides the lambda score in the log output from PLINK.

```
plink2 --pfile
imp/furtherpc/output/filter/merge/mnctrlallchrstatus --glm --
adjust --out imp/furtherpc/output/assoc/allchr/mnctrl
```

6.2.6. Whole genome imputation

Imputation is a technique widely adopted to increase the density of SNV markers from genotyping. Imputation uses information from a reference panel to match shared haplotypes in the dataset of interest and predicts the untyped genotypes, see 5.4.2.2 for more detail. Beagle 5.1 was used for whole genome imputation. Imputation is entirely dependent on the quality of the reference panel.

Imputation was necessary as following quality control there were only 188,662 intersecting SNVs in the case-control dataset. In the unimputed unmerged dataset there were 672,116 SNVs in the AMN dataset and 751,308 SNVs in the control dataset. As a result of the low intersecting SNVs I decided to impute the case and

control datasets separately so that the pre-imputation target SNV markers would be higher in number and this would potentially improve the quality of imputation. As an additional check I also decided to impute the merged case-control dataset with fewer SNVs. All analyses were done on chromosome 2 in the first instance to determine the best methodology before imputation across all 22 chromosomes.

6.2.6.1. European reference panel

Access was publicly available for 2504 individuals from the 1000 Genomes Project, phase 3 [309]. However, only 503 of these individuals are of European ancestry so first these individuals were extracted from the dataset which had 27,520,389 variant markers. This was important as while some studies report using all ethnicities it is known that the success rate will be higher with a better matched reference panel [312]. It should also be noted that while the European subset has high genotyping rates and low imputation error rates, rare SNVs should be excluded from analysis and imputation with, and of other ancestries should be done with caution [313]. In contrast, other studies suggest that if imputing individuals of European ancestry the overall reference panel sample size is more relevant and a mixed ancestry reference panel can be used [312]. Because of this uncertainty I decided to undertake imputation with both reference panels and then compare the results to determine the difference.

6.2.6.2. Preparing the case-control dataset

The filtered European case-control dataset was used for imputation. This has a low number of intersecting SNVs and so as an additional method to try to improve overall

SNV density I decided to impute the filtered AMN and filtered control datasets separately with a view to intersecting SNVs post imputation.

6.2.6.2.1. Extracting separate chromosomes

Imputation is performed on a single chromosome at a time due to computational power for large datasets with large genotyping data. The case-control dataset was converted to a VCF file and at the same time each separate chromosome was extracted using PLINK2. This is because PLINK2 processing is faster than VCFtools which can also be utilised to extract separate chromosomes. The job was batch processed by writing a shell script for all 22 chromosomes. To further facilitate faster analysis times the VCF files were compressed using bgzip (block compression tool).

```
plink2 --bfile imp/furtherpc/preimp/mnctrl --chr 1 --export
vcf --out imp/furtherpc/preimp/mnctrl1
```

```
bgzip -c imp/furtherpc/preimp/mnctrl1.vcf >
imp/furtherpc/preimp/mnctrl1.vcf.gz
```

6.2.6.2.2. Conform-gt

To correctly impute data Beagle requires the genotype file and the reference panel to have matching strand orientation and matching major and minor alleles. To enable this the authors of Beagle have developed and provide access to a program called Conform-gt [314]. Conform-gt adjusts the VCF file so that the chromosome strand and allele order match the reference panel VCF file. It requires the program to run with java, the reference panel, the target genotype VCF file, the chromosome number and the location and name of the output.

```
java -jar imp/conform-gt.24May16.cee.jar \
ref=imp/conformout/ctrl_chr1.conform.vcf.gz \
gt=imp/sourcefile/mnctrl_preimp_chr1.vcf.gz \
chrom=2 \
out=imp/conformout/mnctrl_chr1.conform
```

6.2.6.3. Imputation in Beagle

The output file from conform-gt is ready to use directly in Beagle 5.1. The command for imputation in Beagle consists of multiple different parts, an example for chromosome 2:

```
java -Xmx200g -jar beagle.28Jun21.220.jar \
window=30.0 \
gt=conformout/mnctrl_chr2.conform.vcf.gz \
ref=eurreffile/chr2eur.vcf.gz \
out=output/mnctrlimputed
```

Breaking this down into the separate components makes the understanding of this easier. Firstly, ‘java’ calls the java programming language. ‘-Xmx200g’ tells Beagle that it is allowed to use 200Gb of memory for its processing this can be adjusted dependent on the equipment in use. ‘-jar’ tells the computer the location of beagle so that java can utilise beagle for the rest of the command script. ‘window’ is the size of the sliding window that beagle breaks down the genotyping in to, I used 300,000bp but this can be increased or decreased depending on memory power and the resolution of data required. ‘gt=’ tells Beagle the location of your dataset that you wish to impute. ‘ref=’ tells Beagle the location of the reference file and ‘out=’ tells beagle the location and name for the output file.

This was done for all 22 autosomal chromosomes and for each chromosome two output files were created by Beagle. The first is an output log that summarises Beagle version, run time and the arguments in use and the second is the compressed VCF file with the imputed dataset. Valuable additional information is present in the VCF info field; 'DR2' is the estimated squared correlation between the estimated allele dose and the true allele dose, 'AF' is estimated alternate allele frequency in the samples and 'IMP' to highlight if the SNV has been imputed (this helps differentiate genotyped and imputed SNVs).

6.2.6.4. Quality control post imputation

Quality control is necessary following imputation to avoid false positives and more importantly to exclude incorrectly imputed SNVs and multi-allelic SNVs that are not compatible with a GWAS analysis.

6.2.6.4.1. Imputation quality filtering

The DR2 heading is the dosage R^2 which is a measure of imputation inaccuracy using the highest posterior probability and the true allele dosage [216]. The squared correlation does not depend on the SNV allele frequency and is simply for the sample size and power. Browning *et al.* show that DR2 has a good accuracy when the posterior genotype probabilities are accurately calibrated and informative [216]. There is no universal threshold for exclusion of DR2 dosing, however a decision was made to use a stringent threshold to exclude SNVs with a score <0.8 based on prior research from our laboratory [147]. For the DR2 filtering BCFtools was used with the filter command, as an example:

```
bcftools filter -i 'DR2>=0.8' -Oz
imp/furtherpc/output/mnctrl1.vcf.gz -o
imp/furtherpc/output/filter/dr2/mnctrl1impdr2.vcf.gz
```

Multi-allelic SNVs are present in the imputed dataset and so these had to be filtered out as before to allow for further downstream analysis, this was done using BCFtools and the -m2 and -M2 commands.

```
bcftools view -m2 -M2 -v snps
imp/furtherpc/output/filter/dr2/mnctrl1impdr2.vcf.gz -Oz -o
imp/furtherpc/output/filter/biallele/mnctrl1impbi.vcf.gz
```

The VCF files were converted to PLINK format pfiles. The --snps-only just-acgt tool helped exclude any SNVs with one or more multi-character allele codes that were not one of the four nucleotides; ‘A’, ‘C’, ‘G’ and ‘T’.

```
plink2 --vcf
imp/furtherpc/output/filter/biallele/mnctrl1impbi.vcf.gz --
make-pgen --snps-only just-acgt --out
imp/furtherpc/output/filter/biallele/mnctrl1impbi
```

The per SNV QC was then repeated in PLINK with a call rate <95% being excluded, MAF <1% being excluded and HWE with a p-value <0.0001. I chose less stringent criteria for filtering here to maximise the number of SNVs that would intersect between the cases and controls.

```
plink2 --pfile
imp/furtherpc/output/filter/biallele/mnctrl2impbi --geno 0.05
--maf 0.01 --hwe 0.00001 --make-bed --out
imp/furtherpc/output/filter/merge/mnctrl2filt
```

The filtered QC data for each chromosome needed to be merged to produce a single file for all 22 chromosomes. Then for the cases and controls that had been imputed separately these needed to be merged. The --pmerge tool in PLINK v2.0 was not

functioning hence the creation of a bfile at the last command to enable functionality with PLINK v1.9. The --merge-list tool was used for merging of the chromosomes and --bmerge was used for the cases and control datasets merging.

```
plink --bfile imp/furtherpc/output/filter/merge/mnctrl1filt --
merge-list imp/furtherpc/output/filter/merge/mergechr.txt --
allow-no-sex --make-bed --out
imp/furtherpc/output/filter/merge/mnctrlallchr
```

```
plink --bfile output/filter/plink/mnimpeurfilt --bmerge
output/filter/plink/ctrlimpeurfilt.bed
output/filter/plink/ctrlimpeurfilt.bim
output/filter/plink/ctrlimpeurfilt.fam --allow-no-sex --make-
bed --out output/filter/merge/mn_ctrleur
```

With the merged case-control all chromosome dataset the case and control status needed to be updated in the *.fam file to enable association analyses, at the same time the bfile was converted back to a pfile.

```
plink2 --bfile imp/furtherpc/output/filter/merge/mnctrlallchr
--fam imp/furtherpc/preimp/mnctrl.fam --make-pgen --out
imp/furtherpc/output/filter/merge/mnctrlallchrstatus
```

To ensure that only SNVs with full intersection were kept within the dataset another QC per SNV was conducted.

```
plink2 --bfile output/filter/merge/mn_ctrleur --geno 0.05 --
maf 0.01 --hwe 0.00001 --make-pgen --out
output/filter/merge/mn_ctrleur_filt
```

6.2.7. HLA imputation

SNP2HLA was used to perform HLA imputation which is a plugin to Beagle that adjusts the algorithms specifically for HLA imputation [225]. First an appropriately

genotyped ancestry matched reference panel is required with which to predict the untyped genotypes.

6.2.7.1. European reference panel

The reference panel and sample size is more important than the SNV density for high imputation accuracy [225]. The HapMap European reference panel is included within SNP2HLA and has 124 individuals, with 3924 SNVs and 109 4-digit classical HLA alleles. I applied for access to the larger dataset from the Type 1 Diabetes Genetics Consortium (T1DGC) which contains 5225 European individuals, with 5868 SNVs and 298 4-digit classical HLA alleles [315]. No information was available either on the repository website or in the published article about the disease state of individuals included in the T1DGC reference. On direct questioning the repository support team informed me all control reference panel samples were from parents or siblings of families with an affected individual with type 1 diabetes mellitus. There is a strong association between type 1 diabetes and class II HLA genes with an estimate that they contribute up to 50% of the familial aggregation of disease [316]. The individuals in the reference panel did not have type 1 diabetes however were close relatives of individuals with disease. For this reason, I was uncertain about the use of this as an appropriate reference panel, although it is widely accepted and used in other GWAS and studies as an HLA reference panel [171]. To compare and determine this I decided to undertake HLA imputation with both reference panels so that these results could be compared.

6.2.7.2. Preparing the case-control dataset

The case-control dataset used for HLA imputation is the post ancestry stratification pre-whole genome imputation dataset. This was done to minimise any error with the whole genome imputation being carried forward to HLA imputation. The dataset was in the binary PLINK format.

The cases and controls were imputed separately due to the size of the T1DGC reference panel and computational memory issues the data had to be batched into smaller datasets of 1000 individuals.

6.2.7.3. Imputation in SNP2HLA

An example of the command for SNP2HLA is demonstrated and discussed below:

```
./SNP2HLA.csh mnpreimp HM_CEU_REF 1958BC_IMPUTED plink 2000
1000
```

The first part calls SNP2HLA so that it is used for the subsequent steps. The 'mnpreimp' is the location and name of the desired PLINK bfile format dataset that needs to be imputed, 'HM_CEU…' is the path of the reference panel which contains files in *.bgl.phased and *.markers files. 'plink' tells SNP2HLA the path of PLINK (version 1.0.7), '2000' is the maximum java size in Mb that can be used by Beagle, this can be adjusted accordingly and '1000' is the size of the marker window size. Due the large numbers of individuals in the control dataset the files had to be split into smaller datasets comprising of 1000 individuals each, this was extracted using the awk command in Unix.

6.2.7.4. Quality control post imputation

Like Beagle imputation outputs, SNP2HLA produces a squared correlation score of the quality called, R^2. A stringent threshold to exclude SNVs with a score <0.8 was used. I wanted to assess only the 4-digit HLA types so excluded SNV markers that started with an rsID, the 1kg prefix or had only 2-digit HLA types. Additionally, SNVs with a MAF <1% were excluded.

6.2.8. Association tests

6.2.8.1. Genome wide association test

There are different statistical tests for a GWAS and so selecting the appropriate test and model is important. Statistical tests suitable for analysing binary traits - i.e. the presence or absence of AMN were selected. Based on prior research, data analysis was examined using the prediction of an additive model [73]. In an additive model the addition of an allele increases or decreases the risk of the phenotype in a linear way. This differs from dominant and recessive gene models.

The simplest test is the basic allele test which is based on a 2×3 contingency table for each combination of SNV alleles for cases and controls. Each SNV is tested for association individually with a Chi-square (χ^2) test, see 5.4.2.1 and Equation 6.3. The basic allele test in PLINK is done using the --assoc tool in PLINK 1.9 only:

```
plink --bfile unimputedmn/postpcfilter/eurcasectrl --assoc --
allow-no-sex --out
unimputedmn/postpcfilter/eurcasectrl_assocresults
```

In the recent version of PLINK 2.0 logistic regression analysis is now the standard test for association of binary phenotypes whereby PLINK fits a logistic regression model, Equation 6.4.

$$y = G\beta_G + X\beta_x$$

Equation 6.4: Equation for the calculation of the logistic regression model

Using this equation for every SNV (one at a time) y is the phenotype (case or control status in this analysis), G is the genotype matrix for the current SNV, X is the fixed-covariate matrix. Additional single or multiple covariates can be included [317]. The covariate(s) is predicted to influence the dependent variable and so needs to be corrected for.

```
plink2 --pfile
imp/furtherpc/output/filter/merge/mnctrlallchrstatus --glm --
out imp/furtherpc/output/assoc/allchr/mnctrl
```

For ease of data analysis a p-value filter can be applied to examine the output from the logistic regression analysis using the --pfilter tool.

```
plink2 --pfile
imp/furtherpc/output/filter/merge/mnctrlallchrstatus --glm --
pfilter 0.00001 --out
imp/furtherpc/output/assoc/allchr/mnctrlglm0.00001
```

To correct for fine population stratification principal components can be used as a co-variate file. To create the co-variate principal components the --pca tool is used. This produces the eigenvectors which can then be directly used as an additional covariate in the logistic regression analysis using the --covar tool.

```
plink2 --pfile
imp/furtherpc/output/filter/merge/mnctrlallchrstatus --glm --
pca approx 10 --out
imp/furtherpc/output/assoc/allchr/mnctrlglm_pcs
```

```
plink2 --pfile
imp/furtherpc/output/filter/merge/mnctrlallchrstatus --glm --
covar imp/furtherpc/output/assoc/allchr/mnctrlglm_pcs.eigenvec
--pfilter 0.0001 --out
imp/furtherpc/output/assoc/allchr/mnctrlglmcovar
```

Finally, to test for further independent SNVs within the identified peaks stepwise conditional analyses were conducted using the --condition tool. The lead SNV within the identified signal was then used as a co-variate for the analysis to assess for an independent signal. As there were sequentially further significant SNVs conditional analyses were done on additional SNVs using the --condition-list tool which references a text file with the SNV rsIDs in a list.

```
plink2 --pfile
imp/furtherpc/output/filter/merge/mnctrlallchrstatus --glm  -
condition rs9272532 --pfilter 0.000001 --out
imp/furtherpc/output/assoc/allchr/mnctrlcond1_0.000001
```

```
plink2 --pfile
imp/furtherpc/output/filter/merge/mnctrlallchrstatus --glm --
condition-list imp/furtherpc/output/assoc/allchr/cond2 --
pfilter 0.00001 --out
imp/furtherpc/output/assoc/allchr/mnctrlcond2_0.00001
```

6.2.8.1.1. Clinical parameters

I had access to phenotype information in a small subset of 243 (225 post QC) individuals that were aPLA2Rab positive from a German cohort. Tests for association were conducted with gender, age, aPLA2Rab titres, uPCR, eGFR at presentation and eGFR decline in those individuals with data over 5 years. These analyses were conducted in PLINK; for binary traits logistic regression analyses were done and for quantitative traits linear regression.

6.2.8.2. Multiple testing

The association tests are conducting and analysing over 2 million tests. This results in a large multiple testing burden and can result in detection of false positives. There are different statistical methods to correct for multiple testing in GWAS; Bonferroni correction, Benjamini–Hochberg false discovery rate (FDR), and permutation testing.

The Bonferroni correction is the most widely adopted method and adjusts the p-value threshold by dividing the standard cutoff (0.05) by the number of SNVs being tested. However, this is too conservative because SNVs are in linkage disequilibrium with one another and therefore are not truly independent tests.

Simulations using different multiple testing corrective methods in the WTCC dataset to address this issue were previously investigated [318]. The authors concluded that irrespective of the SNV density the widely used genome wide significance threshold of 5×10^{-8} is adequate at estimating the effective number of tests and therefore controlling for multiple testing in a European population [318].

6.2.8.3. HLA association test

The HLA association test was performed in PLINK 2.0 with the --glm tool. This tool in PLINK v2.0 fits a linear model for quantitative traits and fits a regression model for binary traits. For the HLA association test this was the logistic regression model. Different datasets for the HLA association were examined; all aPLA2Rab positive AMN cases against controls and all anti-THSD7A antibody AMN cases against controls. This was done in the datasets imputed with both the T1DGC reference panel and the HapMap European reference panel. There were 115 HLA types for the

association test so a statistically significant p-value was calculated with a Bonferroni correction. The statistically significant p-value for HLA association testing was calculated as 0.05/115 = 0.00043.

```
plink2 --bfile data/unimpcasectrl/hlaimp/pla2_ctrl --glm --out
data/unimpcasectrl/hlaimp/pla2_ctrl_logassoc
```

6.2.8.4. Epistasis

Epistasis can be checked for with PLINK v1.9. PLINK uses logistic regression for a binary trait and makes a model based on allele dosage (0,1,2) for each SNV. The test in PLINK only considers allelic by allelic epistasis. To reduce computational processing time and as the lead SNVs had been identified, epistasis was checked in the data for the lead SNVs. These were extracted from the original dataset using the --extract tool which reference a text file with the 4 lead SNVs each on a new line.

```
plink2 --pfile
imp/furtherpc/output/filter/merge/mnctrlallchrstatus --extract
imp/furtherpc/output/sigsnv --make-bed --out
imp/furtherpc/output/filter/merge/mnctrlsigsnv
```

Epistasis was then checked for using the --epistasis tool. The default for PLINK is to only output significant data with a p-value <0.0001 so to change this default the --epi1 tool is used.

```
plink --bfile imp/furtherpc/output/filter/merge/mnctrlsigsnv -
-epistasis --epi1 1 --allow-no-sex --out
imp/furtherpc/output/assoc/allchr/mnctrlsigsnvepi
```

Epistasis in PLINK can also be checked for only in cases:

```
plink --bfile imp/furtherpc/output/filter/merge/mnctrl2sigsnv
--fast-epistasis --case-only --allow-no-sex --out
imp/furtherpc/output/assoc/allchr/mnctrl2sigsnvepicase
```

An alternative statistical method based on the W-test was utilised as an additional check for epistasis [319, 320]. The W-test measures the association between a binary phenotype and creates a contingency table with the different number of combinations that are possible with the genotype data (9 for 2 SNVs; 00,01,10,11,12,21,22,02,20). The statistic tests to see if there is a distributional difference between cases and controls through a combined log odds ratio. It is more powerful in lower frequency variants than alternative methods such as logistic regression, Chi-squared test and multifactor dimensionality reduction [319].

This was run in R with a package called 'wtest' [320]. Data were first converted to two separate *.csv files. One with the phenotype status of each individual coded as 1 for cases and 0 for controls, this was transposed so that the format matched as that required for wtest; so that each separate individual was in a different column. For the genotype data this was converted to an additive genotype *.ped file in PLINK which was transposed in to a new *.csv to create a matrix with genotypes in columns and subjects in the rows.

```
plink --file imp/furtherpc/output/filter/merge/mnctrlsigsnv --
recodeA --allow-no-sex --out
imp/furtherpc/output/filter/merge/mnctrlsigsnvrecode
```

Then to process the data in R for the pairwise interaction calculation the following commands are followed.

To load the data:

```
mn.geno <- read.csv("~/Downloads/2leadsnp.csv")

mn.pheno <- read.csv("~/Downloads/mnpht.csv")
```

To undertake the W-test statistics:

```
hf1 <- hf(mn.geno, w.order = 1, B = 1)

hf2 <- hf(mn.geno, w.order = 2, B = 1)

w2 <- wtest(mn.geno, y = mn.pheno, w.order = 2, input.pval =
NULL, input.poolsize = NULL, output.pval = NULL, hf1 = hf1,
hf2 = hf2)
```

To produce a histogram graph:

```
w.diagnosis(data = mn.geno, w.order = 2, hf2 = hf2, main =
NULL, xlab = NULL, ylab = NULL)
```

To calculate the odds ratio for interaction:

```
y <- as.numeric(mn.pheno)

or.snv1.snv2 <- odds.ratio(mn.geno, y, w.order = 2,
which.marker = c(1,2))
```

6.3. Genetic risk score

The GRS is calculated from a set of independent risk variants associated with a particular disease or phenotype based on a prior GWAS. For each individual, the number of risk alleles at each variant is summed (0,1,2) and is weighted by its effect size [227]. The summation of the scores assumes that each SNV has an additive independent risk and so independent SNVs need to be used for the calculation, see 5.4.2.4 for further detail.

$$AMN\ GRS = \frac{no\ of\ rs4664308\ risk\ alleles \times \ln(2.22) + no\ of\ rs2187668\ risk\ alleles \times \ln(6.07)}{4}$$

Equation 6.5: Equation for the AMN genetic risk score (GRS).

A weighted genetic risk score (GRS) was calculated using the previously reported odds ratios for at rs4664308 (*PLA2R1*), and rs2187668 (*HLA-DQA1)* ascertained from an independent study of AMN in Europeans, Equation 6.5 [73]. The calculation was undertaken for each individual. Allele counts of the risk variants were also examined. With a stringent Bonferroni correction the p-value threshold for significance of <0.0033 (0.05 / 15) was used for the GRS.

6.3.1. Case and control selection

DNA from 1409 individuals with biopsy-proven AMN was available from patients recruited across three European centres, this is the same cohort as 6.2.2. Phenotype data was available on a subset of these cases. In 1130 of these individuals, serum taken within 6 months of diagnosis of AMN was available and aPLA2Rab and anti-

THSD7A antibody titres were measured at a single centre (Medizinische Klinik und Poliklinik III, Universitatsklinikum Hamburg-Eppendorf, Hamburg, Germany). In a smaller subset of 243 German individuals from a single centre, uniformly collected phenotype data were available; age at onset of AMN, sex, uPCR at diagnosis, eGFR decline per year (with a minimum of five year follow up data) and treatment with immunosuppression.

Through our collaborators, I also had access to DNA from a small incredibly rare AMN cohort of 15 non-familial European paediatric cases of dual antibody negative biopsy-proven AMN. Age of onset of disease was less than 16 years of age. This DNA was obtained at a later stage and so these individuals were genotyped on the Illumina Infinium OmniExpress 24 v1.2 microarray beadchip [321]. Quality control was done on these cases as per 6.2.5. The steroid sensitive nephrotic syndrome (SSNS) GRS was also calculated from statistically significant independent lead loci as identified in the GWAS from our group [307, 322]. As I was interested in pairwise interactions between the paediatric group alone I undertook the Wilcoxon rank sums test for comparison.

Another rare cohort of anti-contactin-1 antibody associated AMN and CIDP patients were made available to me. I had access to the DNA for 7 biopsy-proven AMN cases with confirmed anti-contactin-1 antibodies that were negative for aPLA2Rab. These cases were genotyped on the Illumina Infinium Multi-Ethnic Global BeadChip and QC was done on these cases as per 6.2.5. The ancestry of these individuals was not known so principal components analysis was done to obtain a European cohort prior to GRS calculation. As I was interested in pairwise interactions between the

paediatric group, I firstly undertook the Wilcoxon rank sums test for comparison and then subsequently undertook Kruskal-Wallis and Dunn's multiple comparison test.

For the control dataset the same three datasets were used as negative controls as those used in the GWAS (WTCC, Oxford and Illumina). I also decided to use a positive nephrotic control to further demonstrate that the results were specific to AMN; 422 individuals with SSNS were used.

6.3.2. Computational tools

There are different computational tools available for calculating a GRS or otherwise sometimes known as a polygenic risk score.

6.3.2.1. R and Rstudio

The statistical analyses for the lead 2 SNV analysis of GRS was conducted in R language through the Rstudio interface [323]. R was used to undertake regression analyses and figures were produced using the packages ggplot2, qqman and EnvStats [323-326]. Kruskal-Wallis and Dunn's multiple comparison test was performed to assess differences with a stringent Bonferroni correction to consider pairwise comparisons of 6 different groups using a p-value threshold for significance of <0.0033 (0.05 / 6C2). For linear regression a standard p-value <0.05 was used.

6.3.3. Datasets for analysis

Investigation and comparison of binary and quantitative traits with the phenotype data was possible in the AMN cases. I was most interested in investigating antibody

status and had data on aPLA2Rab and anti-THSD7A antibody. Comparisons were made with all groups with both healthy controls and renal disease controls [293, 294, 322, 327, 328]. Renal disease controls were individuals with SSNS that met standard international criteria for nephrotic syndrome and steroid sensitivity was defined as per the international guidelines of standard response to steroid treatment within four weeks of treatment [322]. In a smaller subset of individuals association was examined with other clinical parameters, The GRS was calculated in the paediatric and anti-contactin antibody cohorts.

6.4. UK Biobank

The UK Biobank is large biomedical database that provides access to medical and genetic data of almost 500,000 individuals to accredited researchers. I sought to utilise this resource in the UK population to estimate the population risk of AMN [329].

6.4.1. Data accessibility

Data was accessed and downloaded after approval of the application through the UKBB portal and requires a 32-character MD5 Checksum and a 64-character password [330]. Genotyped data is available in PLINK binary format and was downloaded using the following commands:

```
./ukbgene cal -c22 -v
./ukbgene cal -c22 -v -m
./ukbgene cal -c22 ukb_snp_bim.tar
```

These commands download chromosome 22 *.bed, *.fam and *.bim files respectively. UKBB also provides an imputed dataset which greatly increases the number of SNV markers across the genome. The main limitation is that despite being in a *.bgen format (compressed binary format for typed and imputed data [331]) they are incredibly large making data analysis challenging. The data for all 22 chromosomes was downloaded and is available in our in-house data server.

```
./ukbgene imp -c22 -v
```

The 2 GRS SNVs were within the genotyped data and so the imputed data was not examined.

6.4.2. Quality control

Due to issues with imputed dataset size and computational processing power I decided to first undertake quality control from the genotyped data. From the search tool on the UKBB website the two risk AMN SNVs were within the genotyped dataset and so the imputed dataset did not need to be used for the GRS calculation.

The first step was to merge all chromosomes together, this was done using PLINK and the --merge tool and specifying a text file containing the names of the bfiles.

```
plink --bfile rawdata/ukb_chr1 --merge-list
rawdata/controlallchr.txt --make-bed --out rawdata/ukb_allchr
```

Then standard QC protocols as per 6.2.5 in PLINK were done. The per individual QC was: call rate <90%, heterozygosity rate >3 standard deviations and for IBD relatives were excluded if they were greater than or equal to third degree relatives. Due to the large number of individuals in the UKBB dataset and to improve preservation of the maximum number of individuals for IBD the KING robust kinship estimator in PLINK v2.0 was used. The KING-robust kinship estimator creates a relationship matrix and works independent of allele frequencies and in mixed population datasets [332]. It excludes only one member of each pair of samples with kinship. KING kinship coefficients are scaled so the standard IBD cut-off of 0.1875 is 0.094 in the command.

```
plink2 --pfile ukbb/workingfile/ukbunimppruned --king-cutoff
0.094 --out ukbb/workingfile/ukbunimppruned
```

The per SNV QC was done on call rate <98%, minor allele frequency <99% and HWE <0.001. No information on the X-chromosome was available and as the file format was PLINK non-biallelic SNVs had already been filtered out.

6.4.2.1. Population stratification

The GRS score is only valid in those of European ancestry and so these individuals needed to be extracted. Studies utilising the UKBB dataset do not undertake population stratification and instead use self-reported ancestry information. UKBB provide information on individuals' self-reported ancestry, of which 472,725 are of European ancestry. Bycroft *et al.* analyse the UKBB data for PCA analysis using a Bayesian outlier detection algorithm with three intersecting clusters of the largest number of individuals [333]. This list of 409,728 individuals using self-reported ancestry and genetic data is entitled 'white British ancestry subset' although this was not readily accessible [333].

Visualisation of the genetic spread was done using R (see 0). Exclusion of outliers was better done with the assistance of other specific tools rather than a visual estimation of outliers. The available tools were not able to analyse such a large dataset; SVS due to its user interface was unable to import the dataset. The alternative routinely used method called SmartPCA (undertaking an EIGENSTRAT calculation) was also unable to process a dataset of this size [334]. I therefore decided to use the 472,725 self-reported white European individuals for the dataset.

6.4.3. GRS in UK population

From the filtered dataset of self-reported white European ancestry individuals, the two GRS risk SNVs separately were extracted.

```
plink2 --bfile ukbb/workingfile/ukb_allchr_selfeur --extract
ukbb/pla2snv.txt --make-bed --out
ukbb/workingfile/ukb_selfeur_pla2snv
```

To facilitate a uniform dataset only individuals that had both GRS risk SNVs genotyped were examined and those that either had one SNV or none were excluded using the --keep tool in PLINK.

```
plink2 --bfile ukbb/workingfile/plasnp --keep
ukbb/workingfile/hlasnv.fam --make-bed --out
ukbb/workingfile/plasnv_hlasnvmatch
```

To ensure the correct risk allele was examined the minor allele was changed in PLINK to match that per the GRS calculation using the --reference-allele tool in PLINK which specifies and changes the reference allele as per a text file.

```
plink --bfile ukbb/workingfile/plasnv_hlasnvmatch --reference-
allele ukbb/workingfile/mnrefallele.txt --make-bed --out
ukbb/workingfile/plasnv_hlasnvmatch_a1risk
```

An additive dosage file was created, this tool converts the alleles to a numerical count to represent the numbers of minor alleles.

```
plink --bfile ukbb/workingfile/plasnv_hlasnvmatch_a1risk --
recodeA --out ukbb/workingfile/plasnv_hlasnvmatch_add
```

This data was imported and then analysed in R using R studio.

7. Results

7.1. PLA2R1 intronic variant analysis

7.1.1. Patient characteristics

The study population were from the UK glomerulonephritis (GN) consortium. There were 335 European patients with biopsy-proven AMN. The gender ratio was 2.2:1 with 231 males and 104 females. The mean age at diagnosis was 52.5 years (±13.3). Owing to the historical nature of this cohort antibody testing was not in routine practice so their aPLA2Rab status is unknown.

7.1.2. Computational analysis

7.1.2.1. Overview

From the raw sequencing data after alignment to the reference genome 2203 variants were identified by deCODE. These were filtered for poor and medium quality variants. Variants without a matching bi-allelic control or mismatching alternate control allele were also excluded, totalling a further 13 variants. 949 good quality variants were identified in the *PLA2R1* locus. Of the 949 variants, 482 were known and described in dbSNP v147 and 467 were novel. The association of variants was computed with the p-value significance from the chi-squared result for each variant. There were 9 highly scoring variants with a p-value $\leq 10^{-50}$ and 109 variants with a significant p-value $<5 \times 10^{-8}$. The most strongly associated variant was rs528521365, with a p-value 7.96×10^{-227}, see 7.1.2.4. This variant is intronic and is predicted to be in a TFBS.

I was predominantly interested in the functionality of the associated variants so that a biologically plausible explanation could be determined in the intronic regions for *PLA2R1*. For this reason and also because the lead variant was found to be in a potentially functionally relevant site, I decided to examine all variants with a known function first. This would then enable me to focus the further investigation in to not only the lead variant but also any other strongly associated variants with an ascribed function to them. I then focused the analysis on the lead variant examining any other strongly associated variants that may be in linkage disequilibrium with it.

7.1.2.2. Functional variants

Functionality of all *PLA2R1* variants was computationally examined from UCSC genome browser: 33 are in the coding region and 141 in the regulatory region. Of the 33 coding variants there are: 15 missense (non-synonymous) variants, 15 neutral variants (synonymous) and 3 in either exonic or intronic splice sites, Table 7.1. There were no nonsense variants. 10 missense, 9 synonymous and 1 splice site variants are unique to the cases and are not found in the control population.

Chromosomal position	p-value	Functional annotation	rsID
*160808075	1.82×10^{-11}	Non-synonymous	rs3828323
*160885418	4.50×10^{-9}	Non-synonymous	rs35771982
160836409	5.95×10^{-7}	Non-synonymous	
160797866	3.21×10^{-5}	Synonymous	rs201062324
160885442	5.14×10^{-5}	Non-synonymous	rs3749117
160797510	0.00036	Synonymous	
160797862	0.00036	Synonymous	
160797864	0.00036	Synonymous	
160797597	0.00075	Synonymous	rs61094689
160812310	0.002	Non-synonymous	
160833188	0.011	Splice	rs2715918
160801449	0.022	Non-synonymous	
160801450	0.022	Non-synonymous	
160901517	0.025	Synonymous	rs4665143
160879259	0.043	Non-synonymous	rs33985939
160797748	0.052	Synonymous	
160798224	0.052	Synonymous	
160869875	0.052	Synonymous	
160889543	0.052	Synonymous	rs769505521
160798300	0.052	Non-synonymous	
160889571	0.052	Non-synonymous	
160804072	0.052	Splice	
160811706	0.052	Splice	rs372064390
160798233	0.12	Synonymous	
160798385	0.12	Synonymous	
160901310	0.12	Synonymous	
160801452	0.12	Non-synonymous	
160803350	0.12	Non-synonymous	
160825830	0.12	Non-synonymous	
160869843	0.12	Non-synonymous	
160797848	0.13	Synonymous	rs145983336
160808076	0.75	Synonymous	rs72954858
160901353	0.83	Non-synonymous	rs12327936

Table 7.1: Table of all coding variants sorted by p-value, showing chromosomal position on chromosome 2, annotated function and the rsID. Novel variants do not have an rsID number, variants denoted with * are variants previously described by Coenen *et al.* [72].

7.1.2.3. Functional annotation

7.1.2.3.1. Coding variants

There was a total of 33 coding variants identified in AMN cases, 10 had a p-value <0.05 and only 2 had p-value $<5x10^{-8}$, Table 7.1. The 2 significantly associated coding variants were both missense variants: rs3828323 and rs35771982.

Variant rs3828323 is located on 2:160,808,075 and had an association with AMN, p-value $=1.82x10^{-11}$. It is protein altering, p.Gly1106Ser, in exon 24 of *PLA2R1* which is located within the protein domain of the linker region between C-type lectin domain (CTLD) 6 and 7. This is a tolerated variant with a SIFT (sorting intolerant from tolerant) score [335, 336] of 0.84, and has a high minor allele frequency of 0.56 in European controls, making it a common variant but in the cases it was less frequent and had an allele frequency of 0.35. It has been described in the Korean AMN population with the CC genotype having an odds ratio of 1.36 but without reaching statistical significance [337]. It is also one of the 6 common variants identified by exon sequencing of *PLA2R1* [72].

Variant rs35771982's location is 2:160,885,418 and is associated with AMN with a p-value $=4.5x10^{-9}$. It is a missense variant g.160,885,418 or c.898 G>C in exon 5 of *PLA2R1*, with a predicted protein alteration of p.(His300Asp), Table 7.1. This variant was also one of the common variants demonstrated by Coenen *et al.* and alters amino acids in CTLD-1 [72]. The control minor allele frequency is high at 0.49 which is lower in AMN at 0.33. It has been demonstrated to be associated with AMN although the risk allele has been different in different ancestries - C allele in Koreans

compared to a G allele in Taiwanese and Japanese. This could however be due to ambiguity in the DNA strand that the allele was called from.

7.1.2.3.2. Regulatory variants

There were 141 variants in the dataset annotated to be in TFBS, of which 63 had p-values <0.05, of these 7 were highly associated with p-values $<5x10^{-8}$. This included the lead variant discussed below, 7.1.2.4. These variants were associated with a variety of transcription factors, many of which are associated with inflammatory responses, cell proliferation or apoptosis which may be relevant mechanisms in AMN.

7.1.2.3.3. Rare variants

Only 15 rare variants were detected with an allele frequency <1%. All were novel, 8 were missense and 7 synonymous. These could be private variants as the number of alleles with these are low, ranging from 3 to 5.

7.1.2.4. Lead AMN variant

7.1.2.4.1. Overview

The most strongly associated variant was at position 2:160856173 (rs528521365), $p= 7.96x10^{-227}$. This variant (rs528521365) is in an intronic region between exon 11 and 12 and in UCSC is in a predicted TFBS for CCAAT/enhancer binding protein beta (CEBPB). In AMN cases, allele C has an allele frequency of 98% compared to the population allele frequency of 17%, so is present in 653 of 668 case alleles. By contrast, the alternative allele A is present in 2% of cases and 83% of controls. The SNV score for the lead variant was 13.5 (with the cut-off threshold being <10).

7.1.2.4.2. Linkage disequilibrium

To understand the relationship of the lead variant (rs528521365) with the other variants LocusZoom plots were created from control LD data, Figure 7.1. Variants with a known functional effect from UCSC were filtered to visualise possible associations, this demonstrated the variant is in LD with other variants in predicted TFBSs and a splice site, Figure 7.2. Subdividing this further to assess linkage with only the coding regions demonstrated more clearly the strong linkage disequilibrium with a splice site variant and a synonymous coding variant, Figure 7.3. In all these graphs it is clear to see the missing region near exon 1 that does not have any data. This is because of the failure of a primer in the pooled DNA sequencing and corresponds to the large missing region that is highlighted over exon 1 in Figure 6.1.

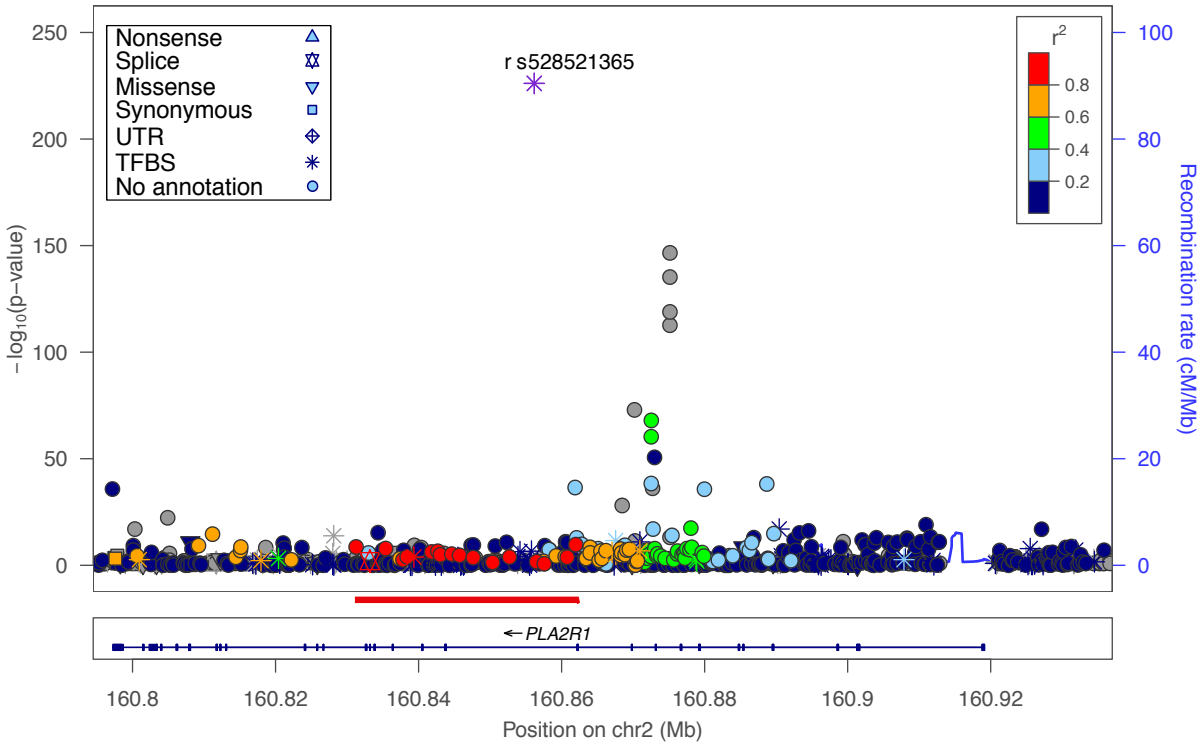


Figure 7.1: LocusZoom plot centred on the lead variant (rs528521365), demonstrating the other variants known to be in linkage disequilibrium with one another in the control data. The key demonstrates the different shapes and the ascribed function. The colour of the r^2 (correlation coefficient) demonstrates the strength of the linkage disequilibrium. The red line highlights the proposed haploblock. The gap near exon 1 corresponds to the failure of a large primer overlying this region.

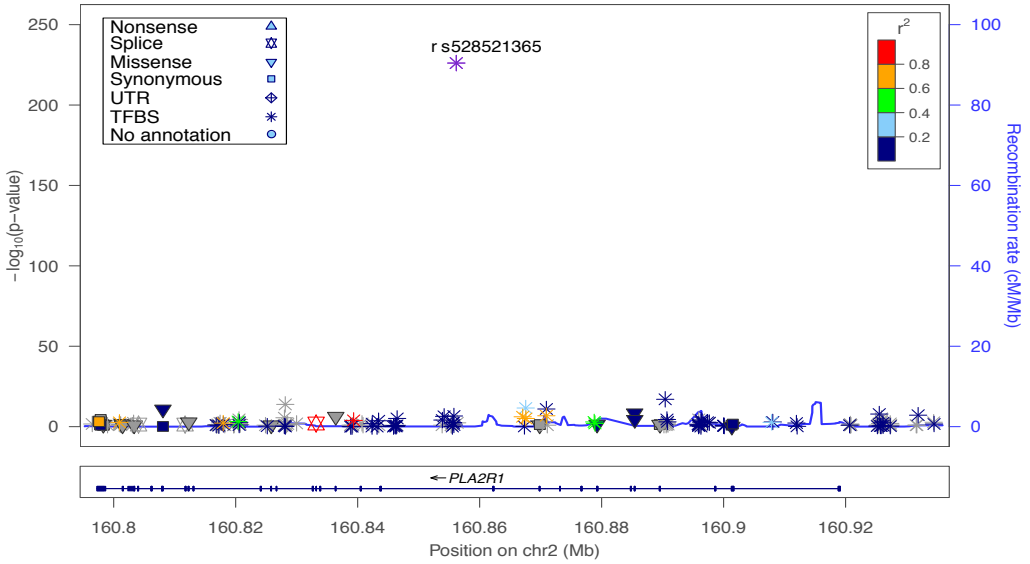


Figure 7.2: LocusZoom plot centred on the lead variant (rs528521365), demonstrating the other variants with an ascribed function on UCSC in linkage disequilibrium with the lead variant. The key demonstrates the different shapes and the ascribed function. The colour of the r^2 (correlation coefficient) demonstrates the strength of the linkage disequilibrium.

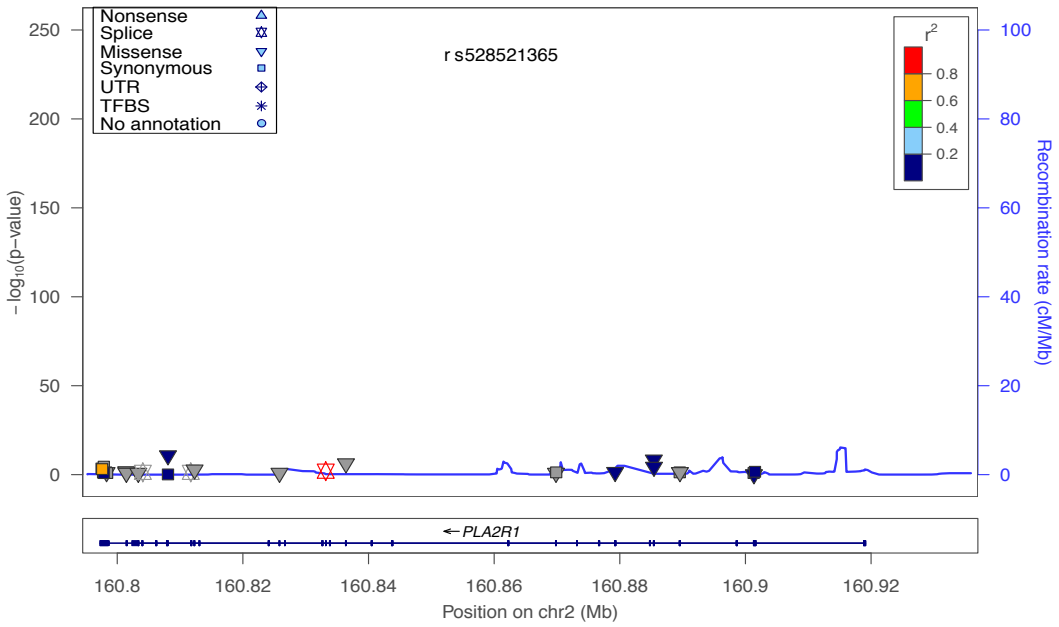


Figure 7.3: LocusZoom plot centred on the lead variant (rs528521365), demonstrating only coding variants with the colour representing the strength of the linkage disequilibrium. Strong linkage disequilibrium can be seen with a splice site and a synonymous variant. The key demonstrates the different shapes and the ascribed function. The colour of the r^2 (correlation coefficient) demonstrates the strength of the linkage disequilibrium.

7.1.2.4.3. Analysis of variants in linkage disequilibrium

Variants in LD with the lead variant were analysed to consider alternative potential associations with AMN that may not have reached statistical significance, Table 7.2. The LD with the splice site variant is very high with an r^2 of 0.99. The splice site is an intronic splice site between exon 16 and 17 and is 8 bp 5' upstream of exon 16 in PLA2R1. The splice site variant (rs2715918) itself does not have a statistically significant association with AMN, p-value =0.01. Further, the splice site variant is common in European controls in the ExAC database. HSF predicts no significant splicing motif alteration with the splice variant. The significance of this is therefore questionable.

The TFBS variant that it is in strong LD (r^2 = 0.99) with the lead variant is rs2667025. This is a predicted TFBS for zinc finger protein 263 (ZNF263). This is a transcriptional repressor and is involved in control of cell growth, cell differentiation and development [338]. This variant is not significantly associated as it does not reach the genome wide level of significance with AMN, p-value =0.0001.

There are four other variants that are in LD with the lead variant, 3 are in TFBS (though none in the predicted DNA binding motifs) and another in the exonic coding region. None of these variants themselves are statistically significantly associated with AMN (statistical threshold for significance is $<5 \times 10^{-8}$). In descending order of LD; the third variant is rs2667011, the fourth variant is rs62176719, the fifth is rs1995951, and the final variant is rs61094689, although there is no information available about how this variant may affect gene expression.

rsID	r^2	Function
rs2715918	0.99	intronic splice site between exon 16 and 17
s2667025	0.99	TFBS for zinc finger protein 263 (ZNF263)
rs2667011	0.74	TFBS for the transcription repressor CCCTC binding factor (CTCF)
rs62176719	0.63	TFBS for TCF7L2 (transcription factor 7 like 2) and FOXA1 (forkhead box A1)
rs1995951	0.62	GATA2 (GATA binding protein 2) and EP300 (E1A binding protein P300)
rs61094689	0.62	3' UTR variant (in exon 30) of *PLA2R1*

Table 7.2: Table of variants in linkage disequilibrium with the lead variant. None of these variants reach statistical significance. rs identifiers (rsID) are shown, with correlation coefficient (r^2) and the ascribed function.

7.1.2.4.4. Functional analysis: CEBPB motif

The CEBPB annotated element from the ChIP-seq region in UCSC for *PLA2R1* covers 468 base pairs but the motif of the actual TFBS is only 14 base pairs and ranges from chr2: 160856239-160856252. The causal motif is ordinarily centrally positioned and this is a method that is used as a confirming diagnostic [339]. The motif for CEBPB is not centrally positioned and instead is directed towards the 3' end suggesting the possibility that there may be a potential mistake in the calling of the binding site motif, Figure 7.4. Utilising an alternative tool to the data from ENCODE ChIP-seq it was confirmed that the CEBPB DNA binding site for this transcription factor was correct. A search on the Motif alignment and search tool (MAST) (MEME suite version 4.12.0) [340] of the 468 base pair CEBPB grey defined UCSC region also predicted this same site [341]. The determined logo (visual representation of the positional weight matrix) on UCSC and MAST are the same, Figure 7.5. Further, this is the same consensus domain sequence from *in vitro* experiments of the DNA binding specificity found by PCR-mediated random site selection of the CEBPB family transcription factors [342].

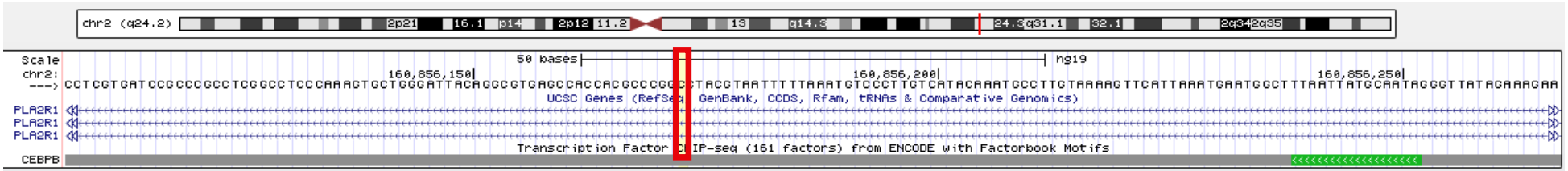


Figure 7.4: Overview of the variant position on the UCSC genome browser hg19. This demonstrates the variant of interest in the red box and its position on the *PLA2R1* gene and the track per which the function was assigned to a transcription factor binding site. The section highlighted green on the CEBPB track is the proposed DNA binding motif for CEBPB.

Figure 7.5: The proposed graphical representation (logo) of the positional weight matrix from UCSC for CEBPB. The identified variant is 66bp downstream of this logo.

7.1.2.4.5. Transcription factor binding sites

7.1.2.4.5.1. TRANSFAC

The TRANSFAC database (version 7.0, Public, 2005) [263] was used to determine alternative potential TFBS near the lead variant chromosomal position. The 468bp TFBS sequence from the CEBPB DNA binding domain from UCSC was scanned using different programs via TRANSFAC. The lead variant of interest in AMN is at position 289 within this sequence. Searching all species there are two potential TFBS for paired box protein (Pax-6) (binding motif starts 4 bp downstream) or vitellogenin gene-binding protein (VBP) (binding motif starts 4 bp downstream of the variant). The reason for this is TRANSFAC uses the same DNA binding motif of CGTAA for both transcription factors. Examining both the variant and the major allele in controls this had no apparent impact on the DNA binding motif as assessed by similarity between the core and matrix similarity scores. The scores for the core similarity score examine just the core part of the DNA binding motif, whereas the matrix similarity score examines the whole matrix for the sequence [343]. Both scores measure the quality of the match between the sequence and the matrix, a score of 1 is an exact match [343]. There was no difference in both scores between the control and the variant sequences. No alternative TFBS were discovered in TRANSFAC.

7.1.2.4.5.2. Patch 1.0

The same search was conducted in Patch [263] which uses sequence pattern recognition to identify potential TFBS [264]. In *Drosophila melanogaster*, the reverse strand has a potential TFBS for ultrabithorax (Ubx, member of the Hox family). With

the lead AMN variant, Ubx would no longer be able to bind to this site. Whilst this is interesting, it is unlikely to have an effect as Ubx is not found in humans [344, 345]. Within the 468 bp sequence Patch also identifies at position 289 the start of a proposed TFBS for c-Fos, c-Jun and CEBPB. This suggests that this is potentially an alternative binding site for a CEBPB complex. With the AMN variant the binding score of CEBPB reduces from 87.5 to 71.4 which may influence the overall binding affinity and efficiency and may be relevant in AMN pathophysiology.

7.1.2.4.5.3. Other prediction tools

There was no difference in TFBS found with the control DNA sequence and AMN DNA sequence in P-Match 1.0 Public [346], MatrixCatch 2.7 [347] and SignalScan [348]. AliBaba 2.1 [349] implicates Sp1 in the region of the lead variant. With the control allele the transcription factor Sp1 binding site does not extend to the position of the variant, however, with the allelic change from an A to a C the Sp1 TFBS extends across this location and the binding site changes from 10 to 14bp.

7.1.2.4.6. Alternative splicing

Analysing the 468bp CEPB sequence surrounding the lead variant in HSF 3.0 [261, 350] did not demonstrate any potential splicing sites near the variant. The variant is at position 289 within the 468bp sequence and the closest splice site is at position 291. However, comparing the control allele and the AMN allele there is no difference in the consensus value suggesting that the variant has no effect on this splice site. HSF also scans MaxEntScan [351] which also identified the same donor splice site but again without any significant alterations with the variant sequence.

7.1.2.5. Other high scoring variants

The next 4 highest scoring variants are within 5 base pairs of each other. This region has a CAAA microsatellite and dbSNP reports a deletion in this region. For this reason, this cluster of 4 variants is in fact a single variant. This has originated due to the sequencing technique and the inability to differentiate copy number variants. Initially they were mistakenly analysed as SNVs by myself until further investigation into the region was undertaken to identify the microsatellite. The association of the supposed SNVs was high. The sixth variant is rs7560040 which has previously been described in dbSNP in association with a deletion. It too is in an area of CT repeats and the allelic variation is described as changing from the reference C to T.

The seventh lead variant is rs2715942, p-value =1.1 $x10^{-68}$. This variant has an allele frequency of T at 69% and C at 31% which is different to the controls (25% and 75% respectively). It is intronic and there is no known function but it is in a DNase hypersensitivity site and is in LD with a TFBS variant.

7.1.2.6. TFBS variants

There are 5 associated predicted TFBS variants to rs2715942 with an r^2 score of >0.6. The strongest LD is with rs3792185 with r^2 =1, the variant itself has a p-value of 0.001 and is associated with CTCF, CBX3 (chromobox 3), RAD21 (RAD21 cohesin complex component) and ZNF143. The next TFBS in LD is rs3792186, r^2 =0.99, p-value =0.01 and this is associated with the same four transcription factors - CTCF, CBX3, RAD21 and ZNF143. The third TFBS in LD is rs2715945, r^2 =0.76, p-value =1.69 x 10^{-7}, and is associated with CTCF. The fourth is rs2715950, r^2 =0.76, p-value =8.1 $x10^{-6}$ associated with EP300 and FOS (fos proto-oncogene). The final

fifth TFBS variant is rs2667011, r^2 =0.74, p-value =7.25 x10^{-7} and is associated with CTCF. In summary, 4 of the 5 TFBS variants in strong LD with the seventh leading variant are associated with CTCF binding from the ENCODE ChIP-seq data. Bioinformatic tools (Patch, P-Match and Matrix Catch) scanning positional weight matrices predict none of these variants are within the DNA binding motif.

7.1.2.7. Haploblocks

The haploblock for the lead variant, rs528521365, extends from 2:160,831,261 to 2:160,879,888, r^2 =0.961 and 0.475 respectively. The strongly associated core of the haploblock (r^2 = 0.8-1) is between 160,830,000 to 160,865,000, Figure 7.6. This extends towards the right (160,865,000 to 160,875,000) to an extended orange haploblock with lower correlation scores (r^2 = 0.6-0.8).

Two different haploblocks were apparent dependent on the variant used for the association plot. The lead variant and the seventh variant were plotted to demonstrate the two strong haploblocks in AMN, Figure 7.6. The first haploblock associated with the lead variant, rs528521365, is between 2:160,830,000 to 160,870,000 and the second haploblock is between 160,870,000 to 160,885,000.

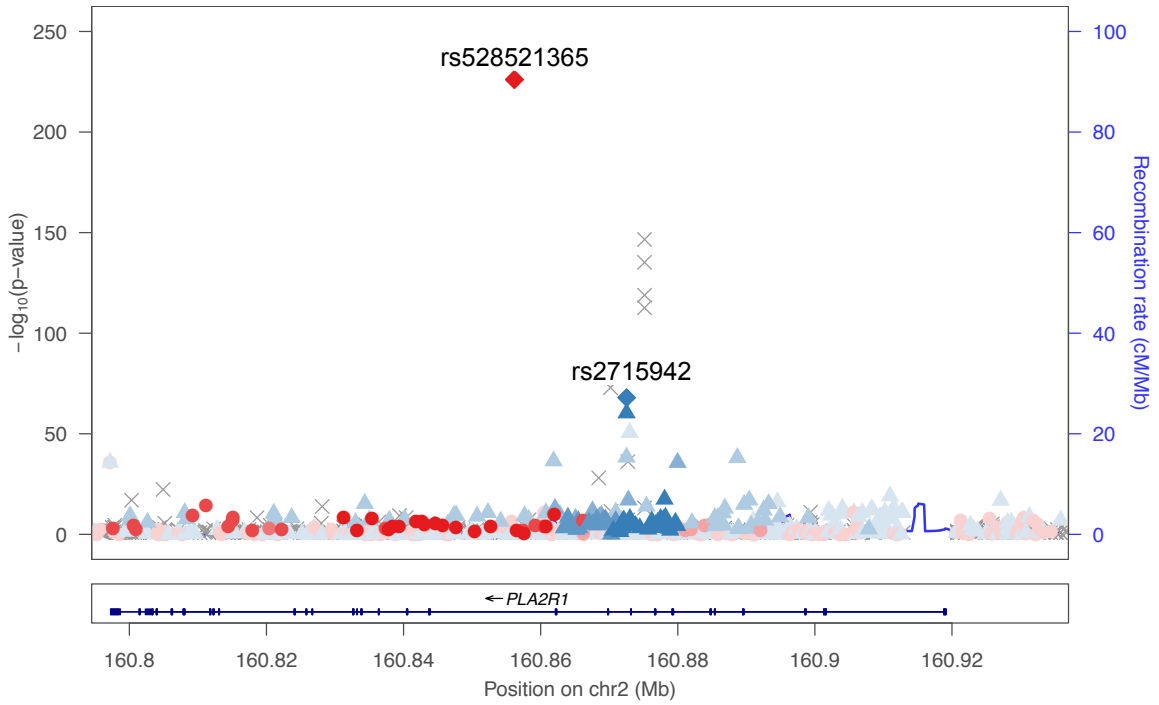


Figure 7.6: LocusZoom plot centred on the two lead variants with available linkage disequilibrium data; this demonstrates the two discrete haploblocks associated with the variants. Red circles represent variants in stronger LD with rs528521365 and blue triangles represent variants in stronger LD with rs2715942.

7.1.3. *In vitro* analysis

There was sufficient computational and statistical evidence that the lead variant (rs528521365) was associated with AMN and that it was associated with the TFBS for CEBPB. No other strong candidates were found that were linked to this lead variant and so I decided to focus the *in vitro* analysis to the lead variant. Of particular interest was that the lead AMN variant may affect the binding of CEBPB and the potential mechanism that CEBPB may have in autoimmune disease, see 8.1.1.1.

7.1.3.1. Transcription and translation of CEBPB

The recombinant CEBPB protein was reconstituted and analysed to ensure that it was the correct product. The results of the Western Blot to visualise the product is

shown in Figure 7.7. Comparison is made to the genomic ladder with known molecular weights. The weight of the reconstituted CEBPB protein was between 37 and 50 kDa (lane 2 in Figure 7.7) and the predicted weight of CEBPB is 39.7 kDa. In the empty vector there were non-specific bands suggesting that other proteins were also present, see lane 3 in Figure 7.7. These same bands were also present in the reconstituted CEBPB product.

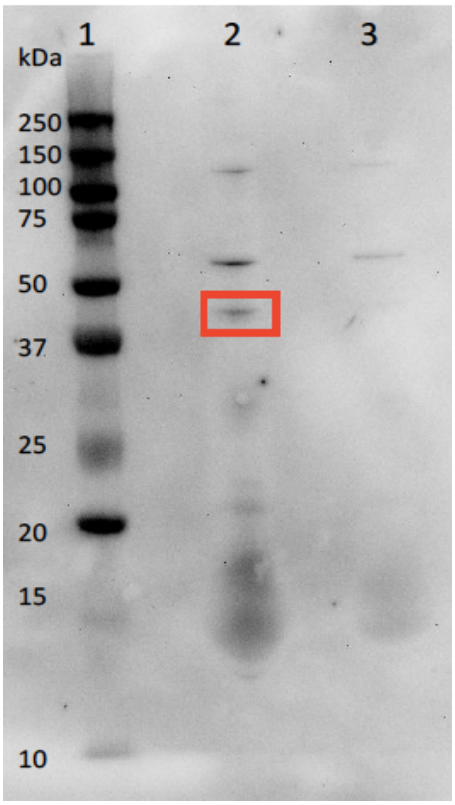


Figure 7.7: Western Blot of production of CEBPB protein from Origene cDNA. The highlighted band is at the predicted molecular weight for CEBPB at 39.71kDa. Lane 1 is the molecular weight protein ladder, lane 2 is the visualised CEBPB band and lane 3 is the control demonstrating no CEBPB band but two other larger bands are visible.

7.1.3.2. EMSA

The functionality and effect of the variant on CEBPB transcription factor binding was assessed with an EMSA. There was no shift or difference between the control and the variant sequence on three separate occasions, Figure 7.8.

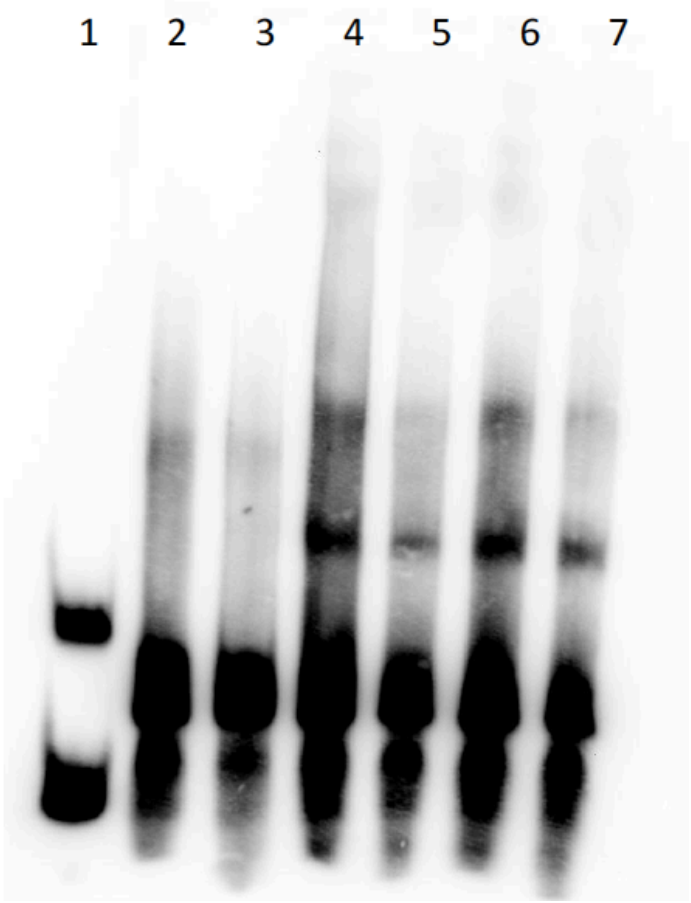


Figure 7.8: Electrophoretic mobility shift assay with and without the lead variant with the transcription factor CEBPB. No shift is seen.
Lane; 1. Positive control: EBV extract with EBV DNA; 2. No protein with variant IMN DNA; 3. No protein with control DNA; 4. TNT proteins & recombinant CEBPB with variant DNA; TNT proteins & recombinant CEBPB with control DNA; 6. TNT proteins (empty vector) with variant MN DNA; 7. TNT proteins (empty vector) with control DNA

7.1.4. Replication

Because no functionality of the lead variant was identified and the SNV score of the variant was low it was decided that replicating these findings in a different cohort with

alternative methodology would be a better way to proceed and save further resources and time. Initially KASP genotyping was outsourced to LGC Genomics. This was unsuccessful so Sanger sequencing was then tried with the goal of sequencing not just the lead variant but also the next cluster of variants 2 to 5.

7.1.4.1. KASP Genotyping

KASP genotyping was performed with the reference and variant sequence for the lead variant. Genotyping was unsuccessful as LGC were unable to make the assay work due to the polymorphic nature and high GC content of the region. The allele that was amplified in their control sample was C/C instead of the major control allele of A/A. LGC tried two different primer designs on 44 samples as part of a validation project and over 50% of the samples stayed in the origin, Figure 7.9A. This compares to the expected results and separation of alleles as undertaken on a validation control sample by myself, Figure 7.9B.

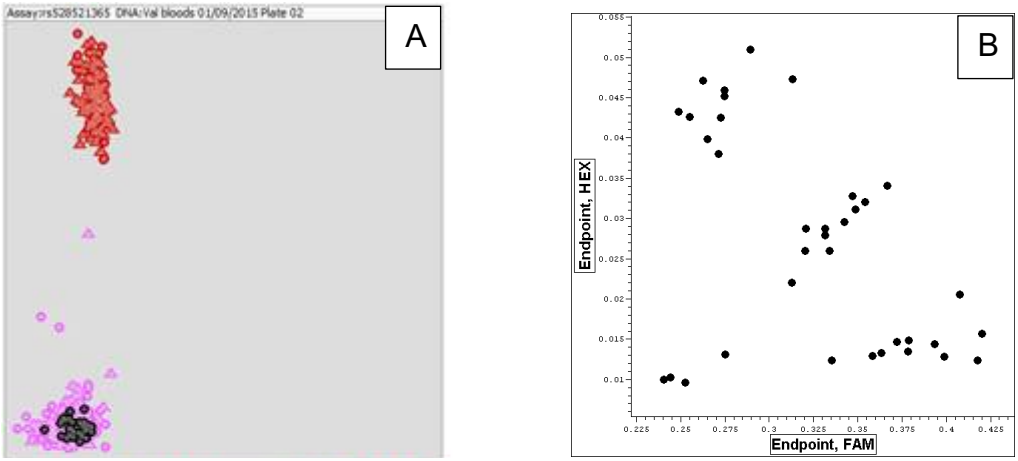


Figure 7.9: A. Results of KASP genotyping control assay with allele specific primers for the lead AMN variant. B. Results of a KASP genotyping control validation assay demonstrating the expected separation of homozygous and heterozygous alleles.

7.1.4.2. Sanger Sequencing

The two regions of interest (lead AMN variant and cluster of variants number 2 to 5) were chosen for Sanger sequencing, Figure 7.10. This first needed to be optimised in healthy control DNA before repeating in the AMN cases.

7.1.4.2.1. Polymerase Chain Reaction

Before Sanger sequencing was possible, I needed to amplify the DNA region of interest in control DNA. 2 sets of primers were selected for each variant that had produced the correct sized product.

For variant 1 the correct bands were 636bp with the primers 1.1 forward and 2.1 reverse and 161bp with 3.1 forward and 3.1 reverse, Figure 7.11. For variant 2 the correct bands were all 455bp with primers 8.2 forward and reverse and 10.2 forward and reverse, Figure 7.12.

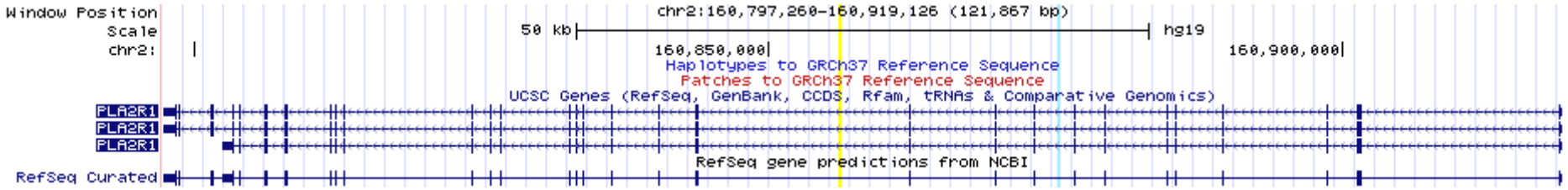


Figure 7.10: Genomic location of the lead variant in yellow and the cluster of variants 2-5 in blue.

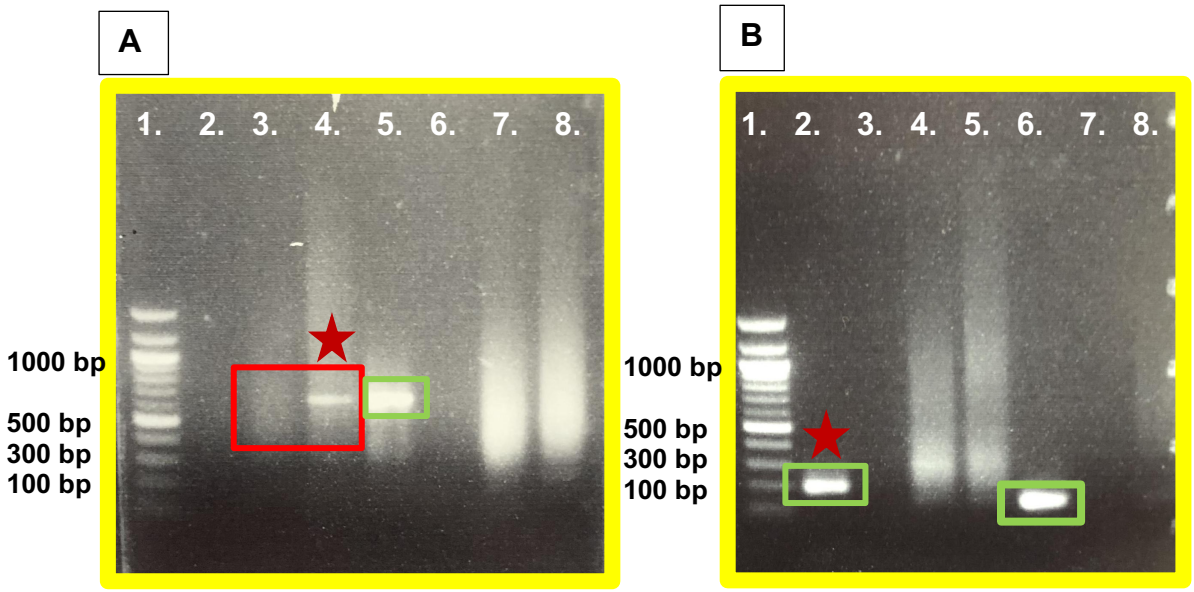


Figure 7.11: Electrophoresis gel UV transillumination result of PCR products for variant 1 (yellow border). The starred boxes were the PCR products sent for sequencing. The box highlights if a correctly sized product was made and the colour of the box represents if the original whole genomic DNA was used (red box) or if nested PCR was used and the amplified DNA was used (green box).

A:
Lane 1 – genomic ladder
Lane 2 – primer 1.1 forward and 2.1 reverse only, no DNA
Lane 3 – primer 1.1 forward and 2.1 reverse & DNA control 1
Lane 4 – primer 1.1 forward and 2.1 reverse & DNA control 2
Lane 5 – primer 1.1 forward and 2.1 reverse & amplified DNA control 2
Lane 6 – primer 3.1 forward and 3.1 reverse only, no DNA
Lane 7 – primer 3.1 forward and 3.1 reverse & DNA control 1
Lane 8 – primer 3.1 forward and 3.1 reverse & DNA control 2

B:
Lane 1 – genomic ladder
Lane 2 – primer 3.1 forward and 3.1 reverse & amplified DNA control 2
Lane 3 – primer 4.1 forward and 4.1 reverse only, no DNA
Lane 4 – primer 4.1 forward and 4.1 reverse & DNA control 1
Lane 5 – primer 4.1 forward and 4.1 reverse & DNA control 2
Lane 6 – primer 4.1 forward and 4.1 reverse & amplified DNA control 2
Lane 7 – primer 5.1 forward and 5.1 reverse only, no DNA
Lane 8 – primer 5.1 forward and 5.1 reverse & DNA control 1

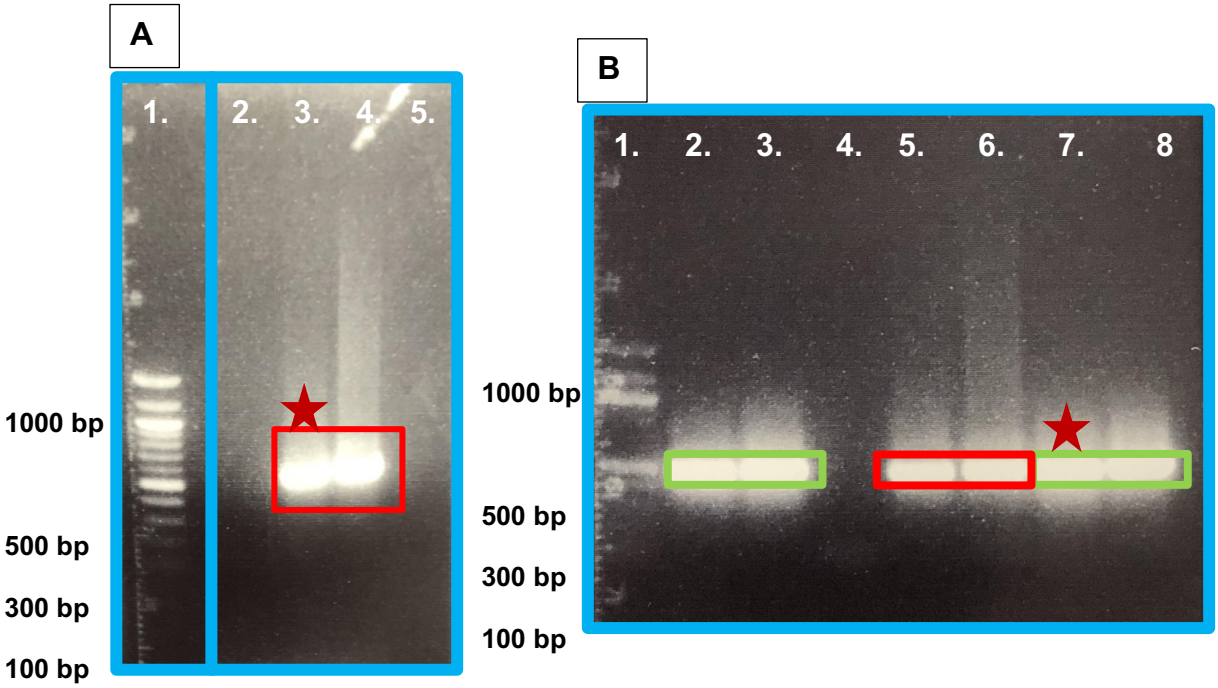


Figure 7.12: Electrophoresis gel UV transillumination result of PCR products for variant 2 (blue border). The starred boxes were the PCR products sent for sequencing. The box highlights if a correctly sized product was made and the colour of the box represents if the original whole genomic DNA was used (red box) or if nested PCR was used and the amplified DNA was used (green box).

A:
Lane 1 – genomic ladder
Lane 2 – primer 8.2 forward and 8.2 reverse only, no DNA
Lane 3 – primer 8.2 forward and 8.2 reverse & DNA control 1
Lane 4 – primer 8.2 forward and 8.2 reverse & DNA control 2
Lane 5 – blank lane

B:
Lane 1 – genomic ladder
Lane 2 – primer 8.2 forward and 8.2 reverse & DNA amplified control 1
Lane 3 – primer 8.2 forward and 8.2 reverse & DNA amplified control 2
Lane 4 – primer 10.2 forward and 10.2 reverse only, no DNA
Lane 5 – primer 10.2 forward and 10.2 reverse & DNA control 1
Lane 6 – primer 10.2 forward and 10.2 reverse & DNA control 2
Lane 7 – primer 10.2 forward and 10.2 reverse & DNA amplified control 1
Lane 8 – primer 10.2 forward and 10.2 reverse & DNA amplified control 2

7.1.4.2.2. Sequencing

The electropherogram demonstrated no signal for either variant with either of the primers, see Figure 7.13 for an example of the results. The sequencing also failed with nested primers on a subsequent attempt, Figure 7.14.

Finally, a combination of nested PCR primers and the pre-amplified control DNA demonstrated some peaks at the third attempt, Figure 7.15 and Figure 7.16. For both variants the expected band length does not correspond to the length of the sequencing and the peaks are all single suggesting homozygosity. This may suggest that only a single allele is being amplified and hence an inaccuracy in the sequencing. Further the quality score is very low for most of the sequencing and is therefore unreliable.

As none of these multiple attempts worked successfully on control DNA samples, I decided to discontinue this work as it was not replicable and therefore not reliable.

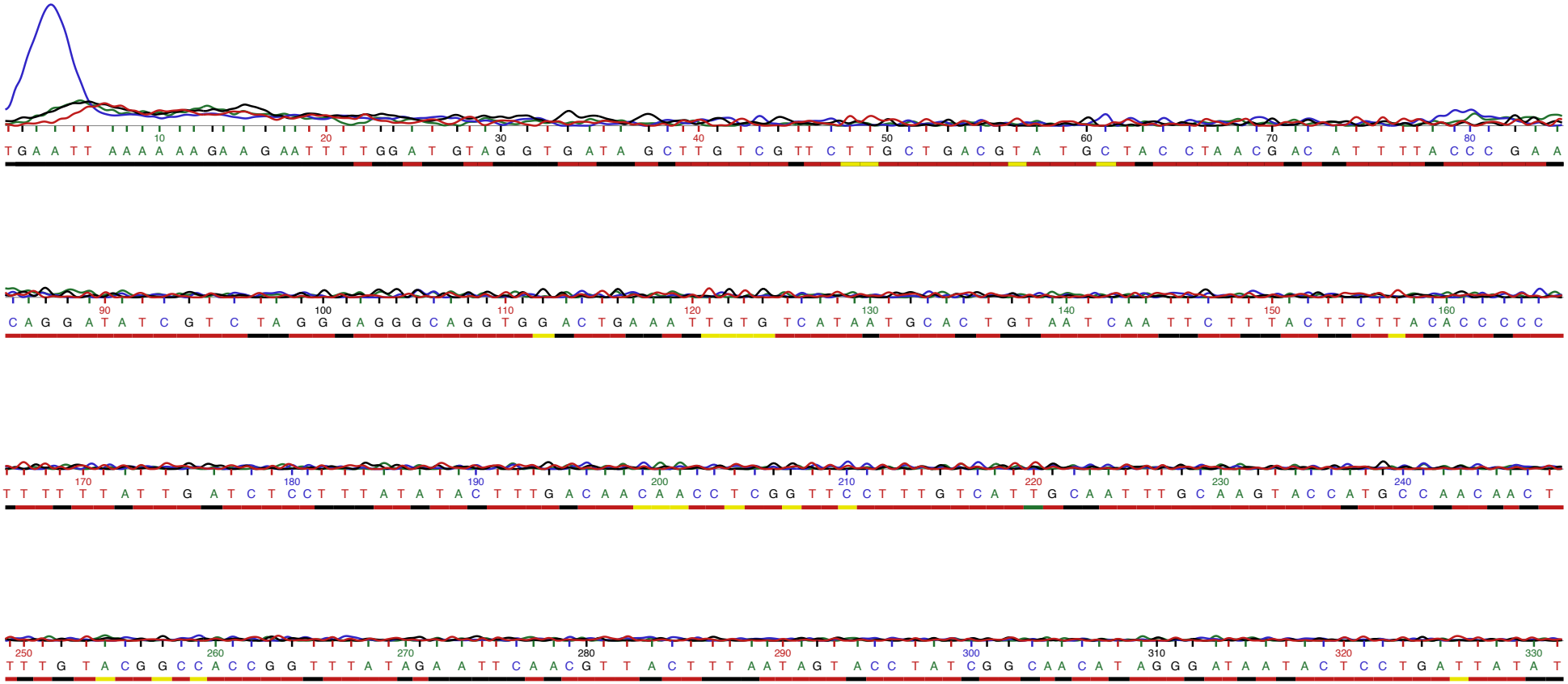


Figure 7.13: Electropherogram of sequencing results for variant 1. No peaks are seen, the sequencing has not worked.

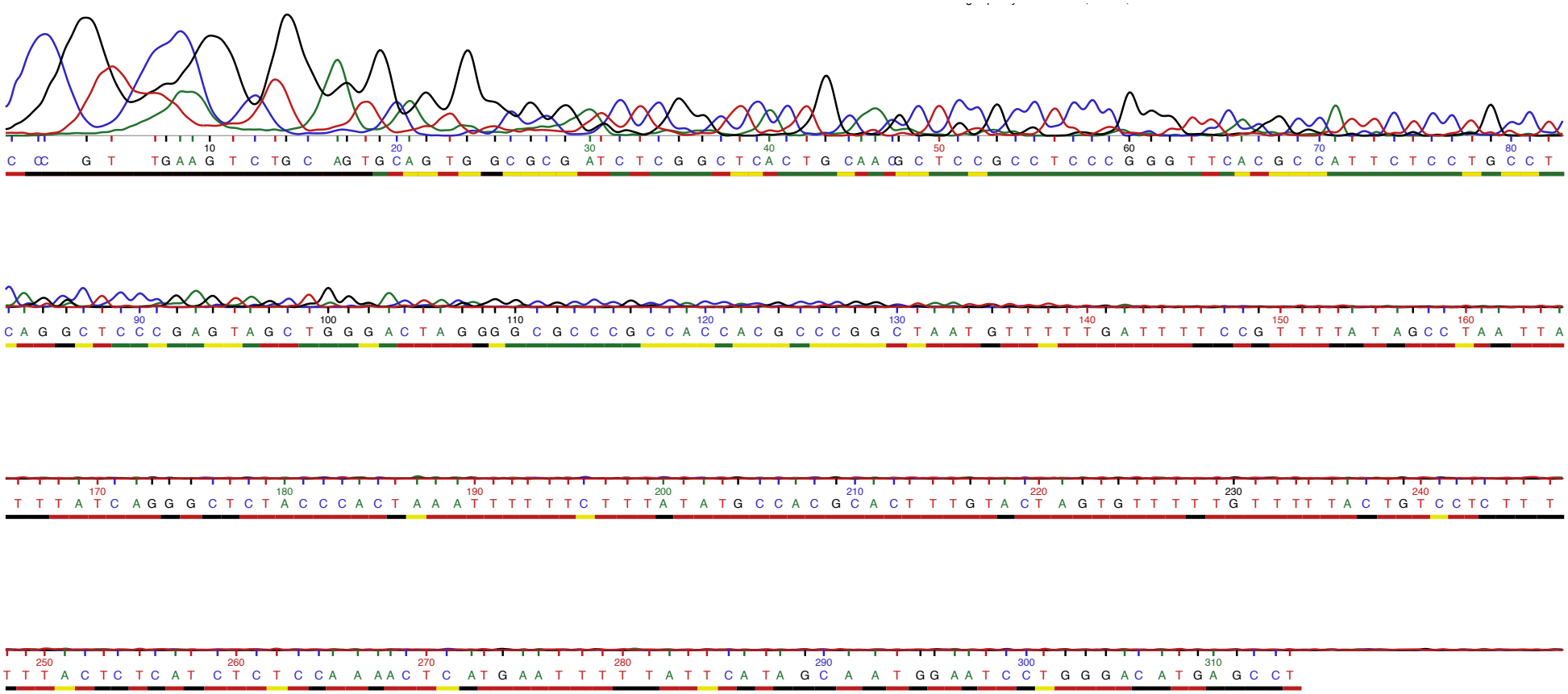


Figure 7.14: Electropherogram of sequencing results for variant 1 with nested PCR primers. Some peaks are seen on the top row but these are inconsistent and do not continue, the sequencing has not worked.

Figure 7.15: Electropherogram of variant 1. Yellow box highlights the SNV of interest to be sequenced. The peaks are not clearly demarcated indicating this has not worked correctly.

A: PCR product from primer 5.1 forward and 6.1 reverse with amplified DNA
B: PCR product from primer 3.1 forward and 3.1 reverse with amplified DNA

Figure 7.16: Electropherogram of variant 1. Blue box highlights the SNV of interest to be sequenced. The peaks are not clearly demarcated indicating this has not worked correctly.

A: PCR product from primer 11.2 forward and 11.2 reverse with amplified DNA
B: PCR product from primer 10.1 forward and 10.2 reverse with amplified DNA

7.2. Genome wide association study

7.2.1. Case dataset

A summary of centres and sample numbers are shown in Table 7.3.

Site	Number
North East & Central London, UK	96
Manchester, UK	917
Hamburg, Germany	420
Total	**1433**

Table 7.3:Table of all AMN individuals genotyped and the source of whole blood samples

A summary of the antibody status subsets are shown in Table 7.17. Clinical parameters of the 225 aPLA2Rab positive cohort are summarised in Table 7.4.

Clinical parameter	Mean / Median	Standard deviation / Interquartile range
Age	55	±15
Gender	Male 0.71: Female 0.39	
Immunosuppression	Received 0.76: None 0.34	
eGFR at diagnosis (ml/min/1.73m^2)	81	49 – 98
eGFR decline (ml/min/1.73m^2/year)	-4.32	-7.7 – 0.41
Urinary protein creatinine ratio at diagnosis (mg/mmol)	6450	4083-10160
Anti-PLA2R1 antibody (Kunits/L)	112	45 – 294

Table 7.4: Clinical parameters of the 225 aPLA2Rab positive cohort (eGFR, estimated glomerular filtration rate; PLA2R1, phospholipase A2 receptor-1)

7.2.1.1. Quality control

7.2.1.1.1. Remedy

Remedy identified 1,748,250 loci from the Illumina microarray bead manifest. There were 940,800 SNVs unmatched to dbSNP. After all the filtering from Remedy there were 754,473 SNVs remaining.

7.2.1.1.2. Ambiguous SNVs

Ambiguous SNVs remained within the dataset and did not allow merging of the data due to strand issues, so these were excluded and filtered out; there were 82,357 ambiguous SNVs.

In the merged dataset for all AMN individuals there were 1409 individuals with 672,116 SNVs.

7.2.1.1.3. Per individual

Quality control of the individuals was done first. Despite sending 1433 individuals for genotyping only 1409 individuals passed the immediate genotyping quality control checks. This could be due to failed genotyping most likely due to insufficient DNA. Post QC (following steps) there were 1269 AMN individuals, see Figure 7.19 for summary.

7.2.1.1.3.1. Call rate

To examine the call rate the number of individuals missing per SNV were reviewed in a histogram, Figure 7.17. From the histogram I decided to use a stringent threshold to exclude any individual with a genotyping call rate <98%. This excluded 32 individuals.

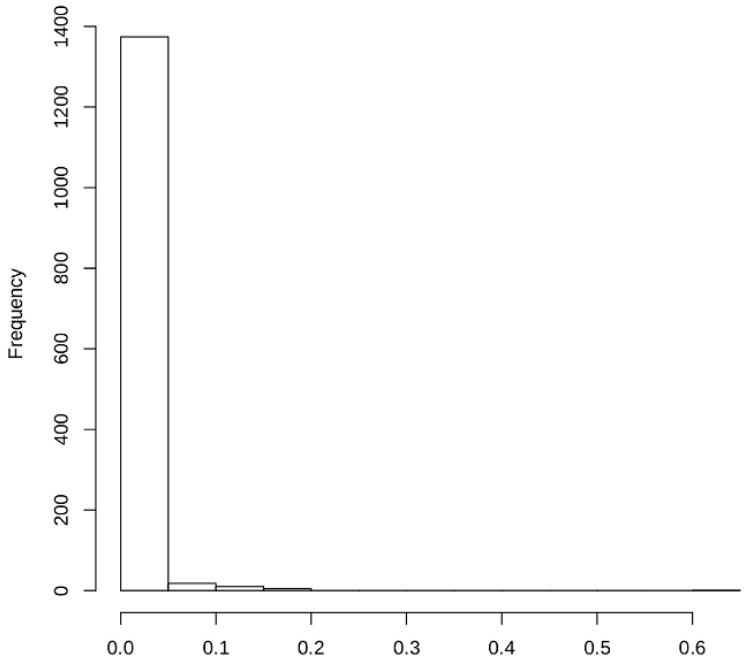


Figure 7.17: Histogram showing proportion of cases with missing SNVs

7.2.1.1.3.2. Heterozygosity rate

The heterozygosity rate was calculated in a pruned dataset of 98,676 independent SNVs and a graphical representation of the proportion was visualised to determine which parameters to exclude, Figure 7.18. A decision was made to exclude individuals ±3 standard deviations from the mean. This removed 57 individuals.

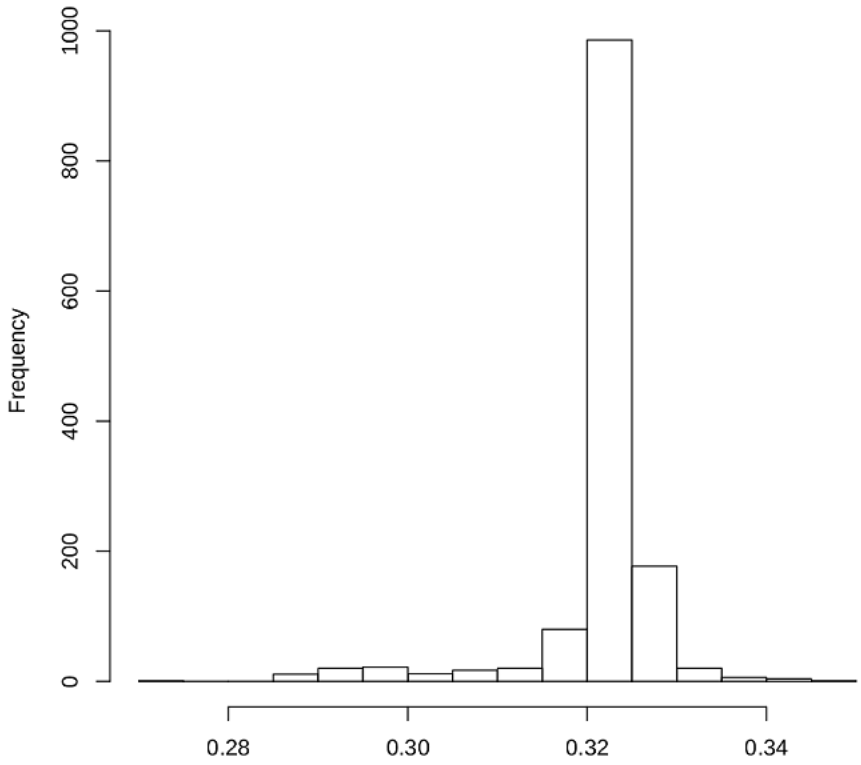


Figure 7.18: Histogram of heterozygosity rate in cases (with 98,676 pruned SNVs)

7.2.1.1.3.3. Identity by descent

IBD filtering >0.1875 excluded 51 individuals.

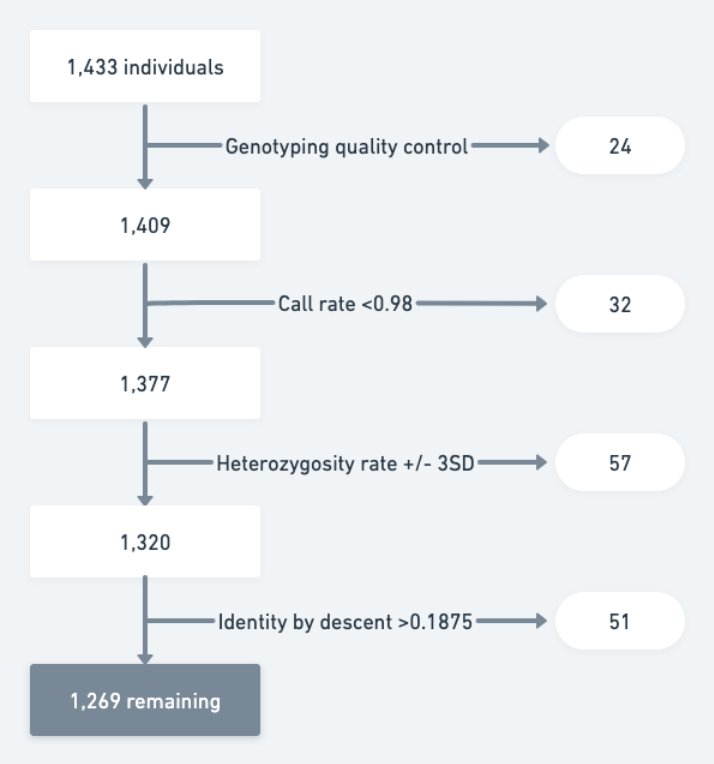


Figure 7.19: Flowchart of per individual filtering in the case dataset (SD = standard deviation)

7.2.1.1.4. Per SNV

Next, quality control per SNV was undertaken. This process had already been started by Remedy which had filtered from 1,748,250 to 672,116 SNV markers. See Figure 7.22 for the per SNV filtering summary.

7.2.1.1.4.1. Call rate

I graphically visualised the spread of the call rate for all SNVs and then using this decided to use a stringent threshold and exclude any SNV with a genotyping call rate of <98%, Figure 7.20. This excluded 43,979 SNVs and left 628,137 SNVs for continuing QC.

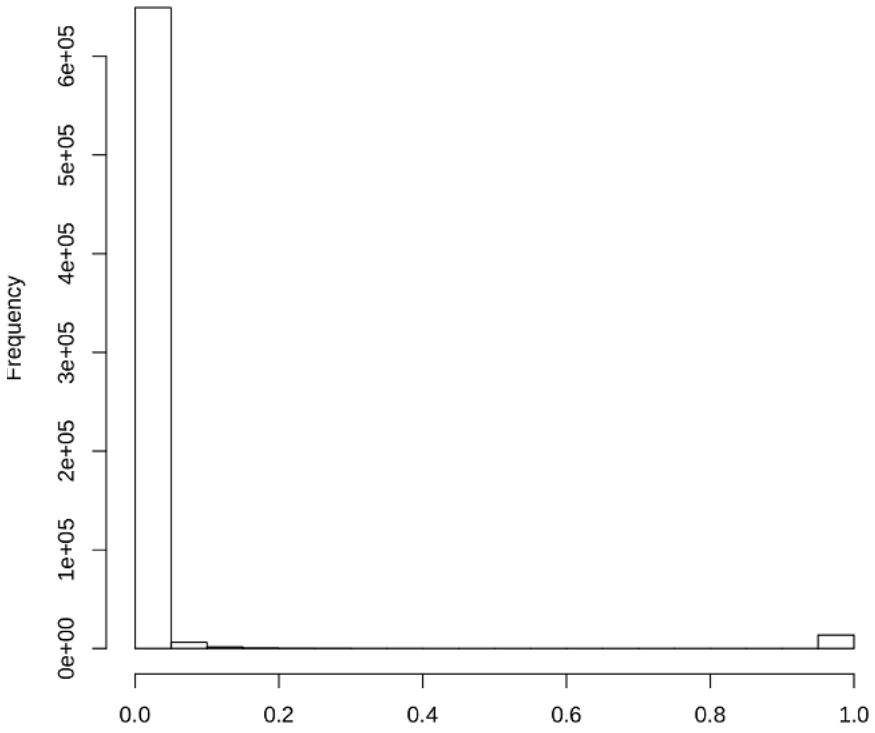


Figure 7.20: Histogram of SNV call rate; showing proportion of SNVs with missing proportion call rate.

7.2.1.1.4.2. Minor allele frequency

A histogram was plotted to visually review the spread of the SNVs MAF, Figure 7.21. Removing SNVs with a MAF of <1% excluded 166,100 however from the histogram to include good quality common variants only I decided to exclude SNVs with a MAF <5%. This excluded 280,863 SNVs and left 347,274 SNVs for downstream analysis.

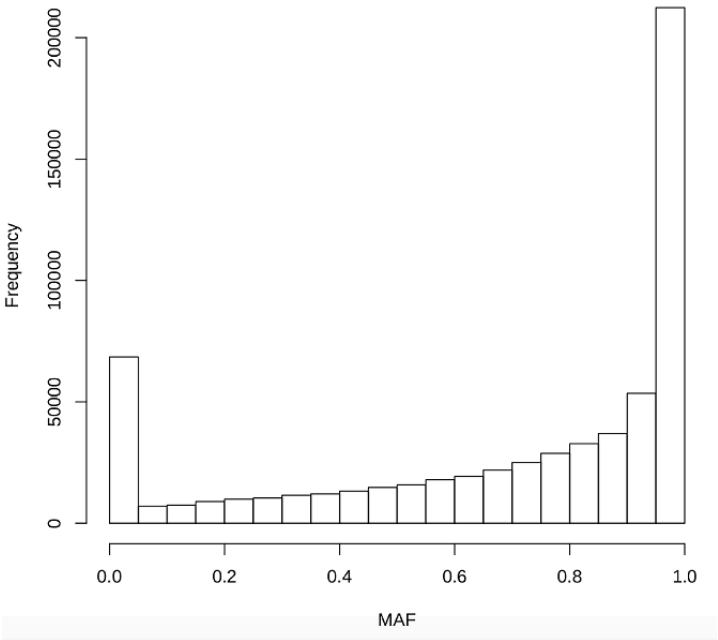


Figure 7.21: Histogram of minor allele frequency proportion of SNVs

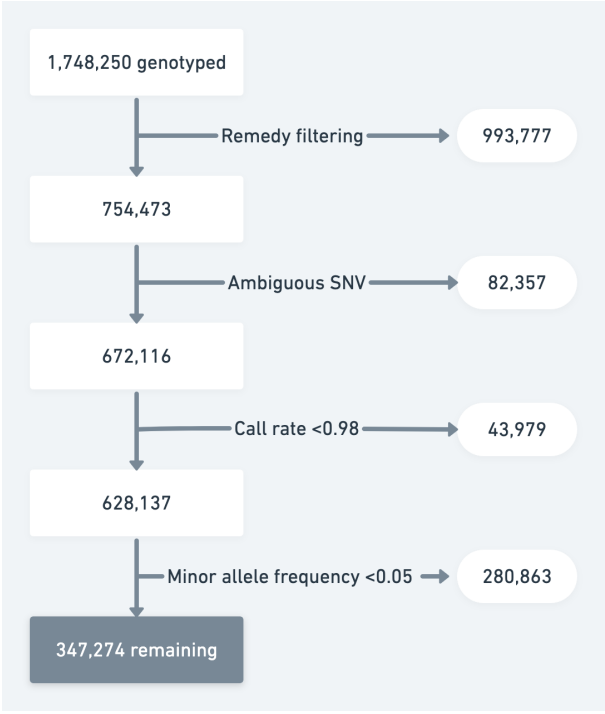


Figure 7.22: Flowchart of filtering per SNV done on the case dataset

7.2.2. Control dataset

Data was downloaded from three different sources of publicly available control datasets; the Oxford cohort, the Illumina cohort, and the Wellcome Trust Case Control Consortium Controls (WTCC). Overall, this gave a combined control dataset of 6036 individuals. These individuals were from European countries and were reported to be European but their ancestry was not confirmed.

The Oxford cohort contains data on 432 self-reported European healthy volunteers; 144 individuals using the HumanOmniExpress-12 v1_J microarray chip with 730,525 SNV; and in 288 individuals on the HumanOmniExpress-12 v1_A with 733,302 SNV markers. Overlapping SNV markers on both microarray chips totalled 730,397.

The Illumina ethnicity control dataset contains 270 individuals across 4 different ethnicities, and for the purposes of population stratification initially I downloaded all data on all individuals. Genotyping is done on a HumanOmniExpress-12 v1_C microarray chip with 731,442 SNV markers.

The WTCC dataset includes; 2,732 self-reported White individuals born in England, Wales and Scotland in 1958; and 2602 healthy blood donors. Genotyping was done on an Illumina 1.2M Duo Custom BeadChip with 1,106,184 SNV markers.

7.2.2.1. Quality control

7.2.2.1.1. Per individual

In total, from 6036 individuals after per individual filtering there were 5257 controls for subsequent analyses, see Figure 7.23 for a summary.

7.2.2.1.1.1. Call rate

Call rate was used to exclude individuals with a high rate of missing genotypes. The overall aim was to be obtain overlapping SNVs as the case dataset so I decided to use a less stringent call rate criteria for the controls. I excluded individuals with a call rate <98%. This excluded 662 individuals and so 5374 were left.

7.2.2.1.1.2. Heterozygosity rate

Filtering for extremes of heterozygosity rates was done. Using the standard criteria to exclude individuals ±3 standard deviations from the mean. This excluded 58 individuals and 5316 were remaining.

7.2.2.1.1.3. Identity by descent

In the control datasets individuals with an IBD score >0.1875 were excluded. 59 individuals were excluded and so 5257 individuals remained passing per individual QC.

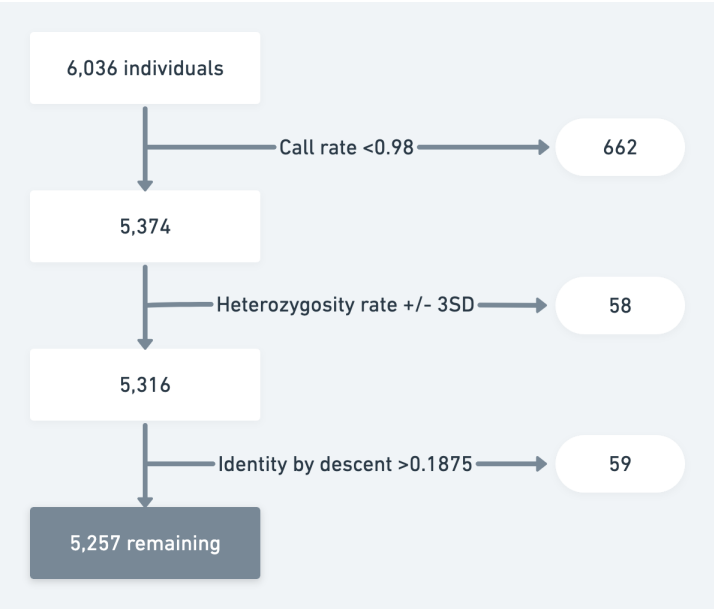


Figure 7.23: Flowchart summary of per individual filtering in the control dataset

7.2.2.1.2. Per SNV

QC per SNV was first undertaken by Remedy excluding 40,469 SNVs. Resulting in an overall 1,065,715 SNVs across the whole control dataset, see Figure 7.24 for a summary.

7.2.2.1.2.1. Call rate

Each SNV with a genotyping call rate <98% was excluded. This excluded 103,402 SNVs leaving 926,313 passing call rate filtering.

7.2.2.1.2.2. Minor allele frequency

SNVS with a MAF <5% were excluded; 285,636 were excluded and 753,644 remained.

7.2.2.1.2.3. Hardy-Weinberg equilibrium

HWE filtering with a high threshold cut-off for SNVs with a p-value <0.001 was used. This excluded; 2,336 SNVs and for subsequent analysis there were 751,308 SNVs.

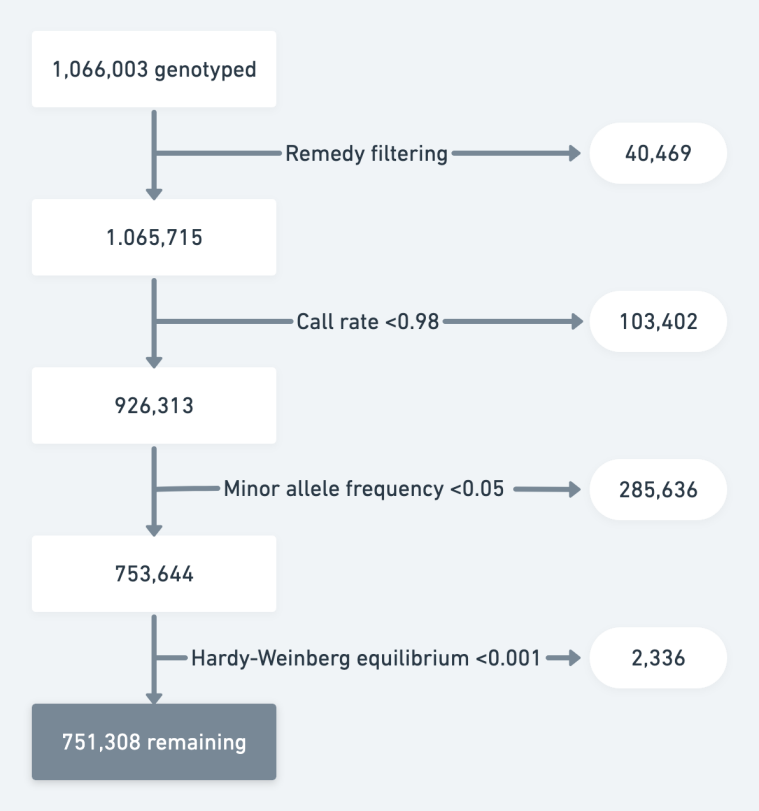


Figure 7.24: Flowchart summary of per individual filtering in the control dataset

7.2.3. Population stratification

The next step for the case and controls was to undertake population stratification to ensure only individuals from a European ancestry remained for analysis, see 6.2.5.2.

7.2.3.1. Principal components analysis

7.2.3.1.1. Illumina ancestry controls

PCA was undertaken using the Illumina ancestry controls, this comprises 270 individuals; 89 African, 91 European and 90 East Asian individuals. The spread of all cases and controls' ancestry is shown in Figure 7.25.

Filtering was done using the standard deviation (SD) from the European ancestry controls. The best scenario was examined by trialling different SD cut-offs. First, a non-stringent SD was tried with 3, Figure 7.26, this excluded 132 AMN cases and 439 controls. Then a more stringent SD cut-off was examined at 2.5; this excluded 143 AMN cases and 1053 controls, Figure 7.27. An even more stringent SD cut-off was tried next at 2.25; this excluded 156 AMN cases and 2441 controls, Figure 7.28.

I decided to strike a balance between excluding too many cases and controls and including too many outliers. For this reason, I chose an SD cut-off of 2.5 and excluded 143 AMN cases and 1053 controls. The final number of remaining AMN cases for analysis was 1125 and controls was 4204.

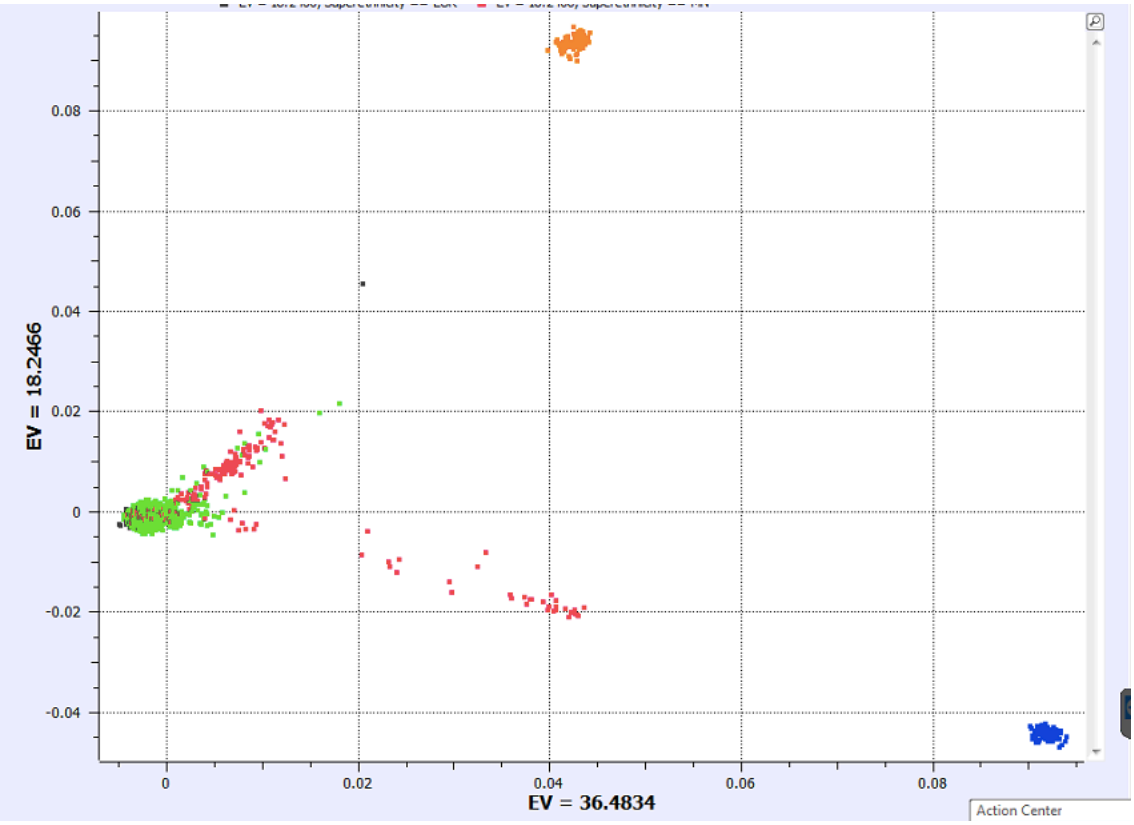


Figure 7.25: Principal component analysis to show divergent ancestries in AMN cases and reference controls. AMN cases are highlighted in red, controls in green, European ancestry controls in black, African ancestry controls in blue and East Asian ancestry controls in orange.

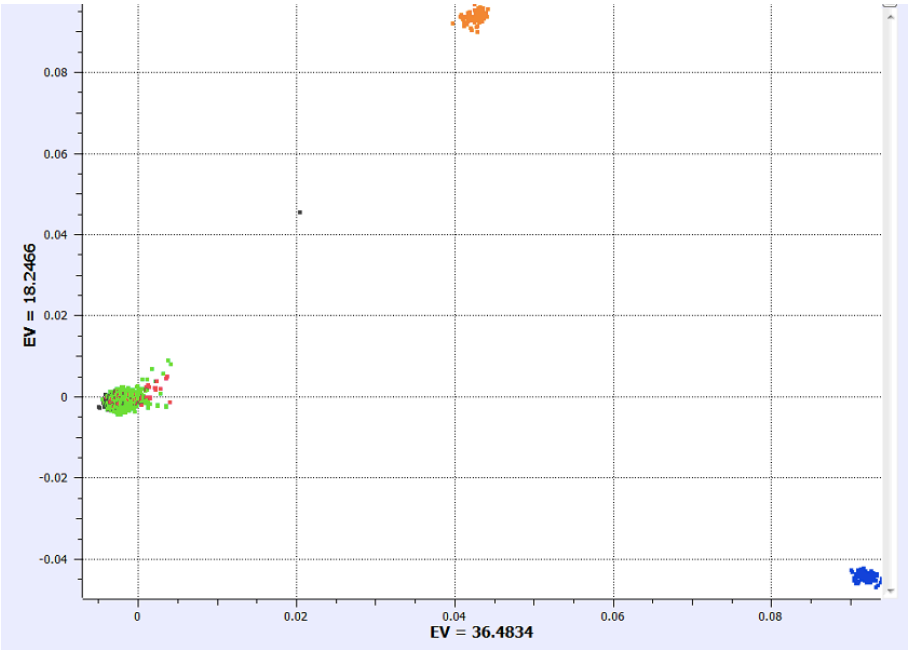


Figure 7.26: Principal component analysis of divergent ancestries in AMN cases and reference controls. Filtered with a standard deviation of 3 from the European ancestry controls. AMN cases are highlighted in red, controls in green, European ancestry controls in black, African ancestry controls in blue and East Asian ancestry controls in orange.

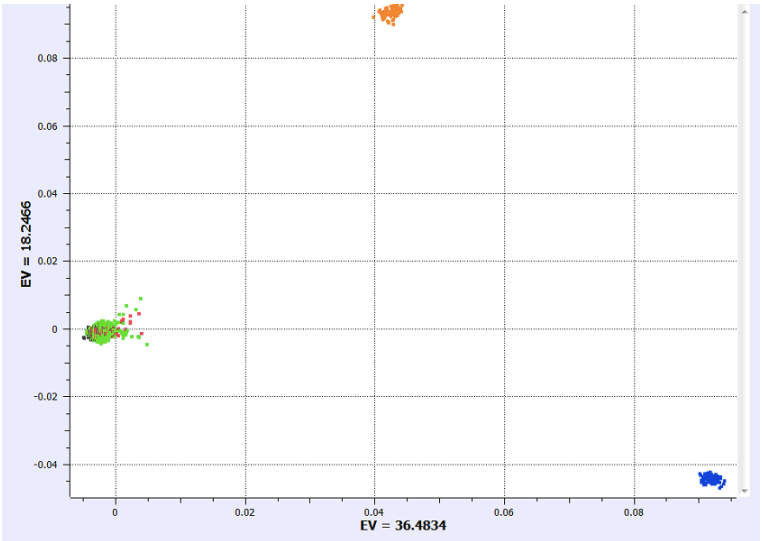


Figure 7.27: Principal component analysis of divergent ancestries in AMN cases and reference controls. Filtered with a standard deviation of 2.5 from the European ancestry controls. AMN cases are highlighted in red, controls in green, European ancestry controls in black, African ancestry controls in blue and East Asian ancestry controls in orange.

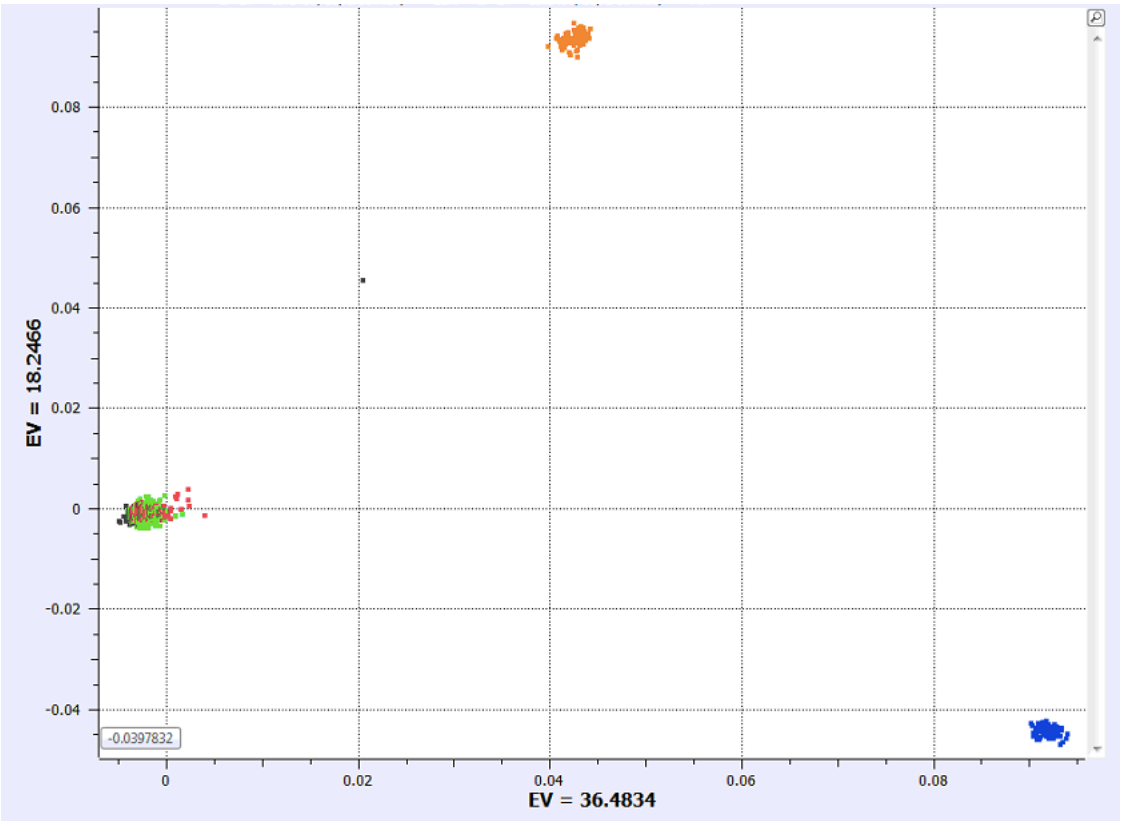


Figure 7.28: Principal component analysis of divergent ancestries in AMN cases and reference controls. Filtered with a standard deviation of 2.25 from the European ancestry controls. AMN cases are highlighted in red, controls in green, European ancestry controls in black, African ancestry controls in blue and East Asian ancestry controls in orange.

7.2.3.1.2. 1000 Genomes Project ancestry controls

Because whole genome imputation was going to be done with the 1000 Genomes Project European individuals as the reference panel, I wanted to ensure there was good overlap with these datasets. As an additional check prior to imputation, I decided to do PCA with the case-control dataset and the 629 ancestral controls. Before intersection of variants between the three datasets (case, control and ancestry controls) some QC was required.

In the 1000 Genomes Project the data was filtered with the standard QC measures, these are summarised in Table 7.5. This left 5,808,310 SNVs in the 629 individuals in the 1000 Genomes Project dataset.

Exclusion criteria	Numbers excluded
Call rate <98%	0 individuals
SNV call rate <98%	17,247,743 SNVs
MAF <5%	2,432,435 SNVs

Table 7.5: Summary table of exclusion criteria on 1000 Genomes project ancestry controls

QC for the combined case-control dataset was also undertaken as an additional check to ensure good overlap and intersection of SNVs, the details of this are summarised in Table 7.6.

Exclusion criteria	Numbers excluded
SNV call rate <98%	720,230 SNVs
MAF <1%	0 SNVs
HWE <0.001	56 SNVs

Table 7.6: Summary table of exclusion criteria on combined case-control dataset

The final number of individuals in the case control dataset was 1269 cases and 5257 controls with a total of 188,662 SNVs.

After merging the three datasets together there were 179,928 SNVs in a total of 7155 individuals (1199 cases, 5249 controls and 629 ancestral controls).

Calculation for PCA were done for the top 10 principal components in an unpruned dataset for all 179,928 SNVs. These were then plotted to visually represent the spread of ancestry in the dataset, Figure 7.29.

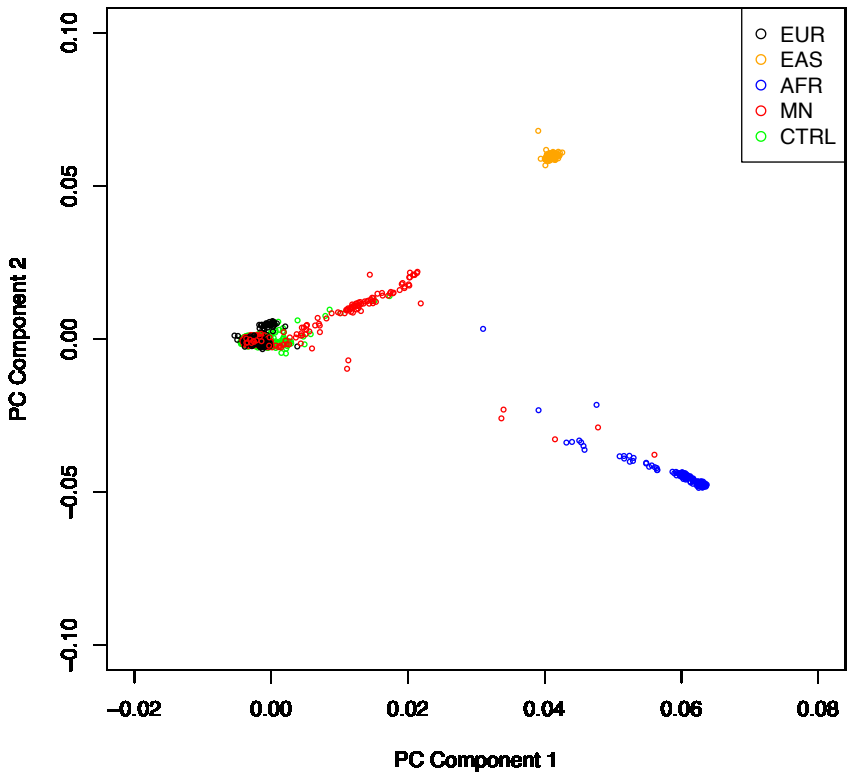


Figure 7.29: Principal component analysis of divergent ancestries in AMN cases and controls with the ancestry controls from 1000 Genomes Project. AMN cases are highlighted in red, controls in green, European ancestry controls in black, African ancestry controls in blue and East Asian ancestry controls in orange.

Different criteria were examined to determine the best SD cutoff, these are summarised in Table 7.7 and shown in Figure 7.31. I decided to use an SD cut-off of 3.25 and excluded 111 AMN cases and 70 controls. This left a dataset of 1158 AMN cases and 5187 controls.

Standard deviation cut-off	No of AMN cases excluded	No of Controls excluded	Corresponding PCA plot
2.0	845	2560	
2.5	348	947	Figure 7.31f
3.0	152	138	Figure 7.31e
3.25	111	70	Figure 7.31d
3.5	103	50	Figure 7.31c
3.75	102	42	
4.0	96	36	Figure 7.31b

Table 7.7: Table of the different standard deviation cut-offs analysed for extraction of European ancestry cases and controls with the 1000 Genomes Project ancestry controls.

7.2.3.1.2.1. Further principal component analysis

After conducting the association test after chromosome 2 imputation there was still considerable population stratification (see 7.2.6.2.2) for this reason I decided to undertake further principal component analysis of the intersecting SNVs in the case and control dataset, Figure 7.30a. This demonstrated continuing population stratification, so I filtered further to reduce this population stratification and chose a standard deviation of 2.5 which removed a further 123 AMN cases and 107 controls resulting in a final number of 1035 AMN cases and 5080 controls, Figure 7.30b.

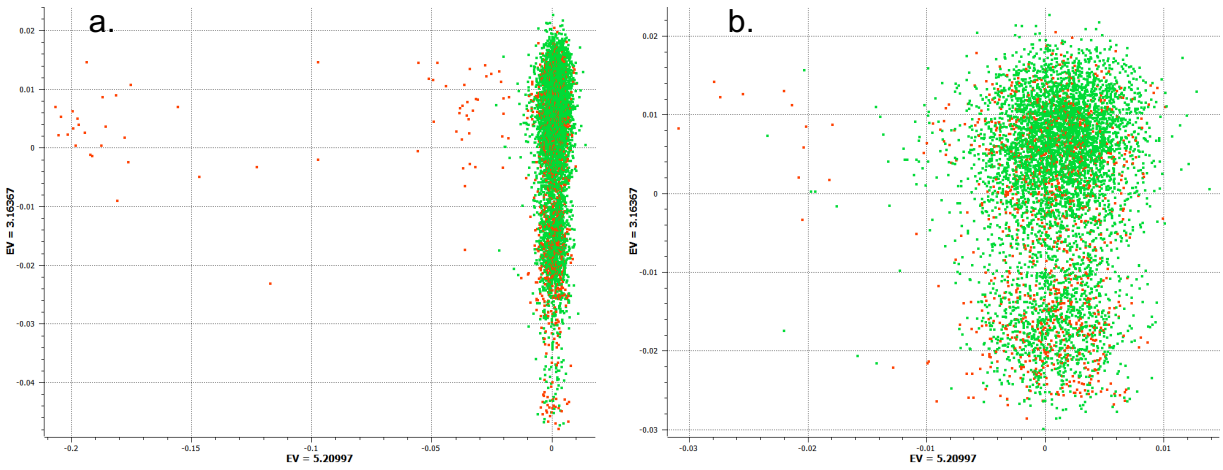


Figure 7.30: Principal component analysis of AMN cases (red) and controls (green). a. Unfiltered principal component analysis with 1158 AMN cases and 5187 controls showing outliers visible to the left. b. Filtered for standard deviation >2.5 resulting in 1035 cases and 5080 controls.

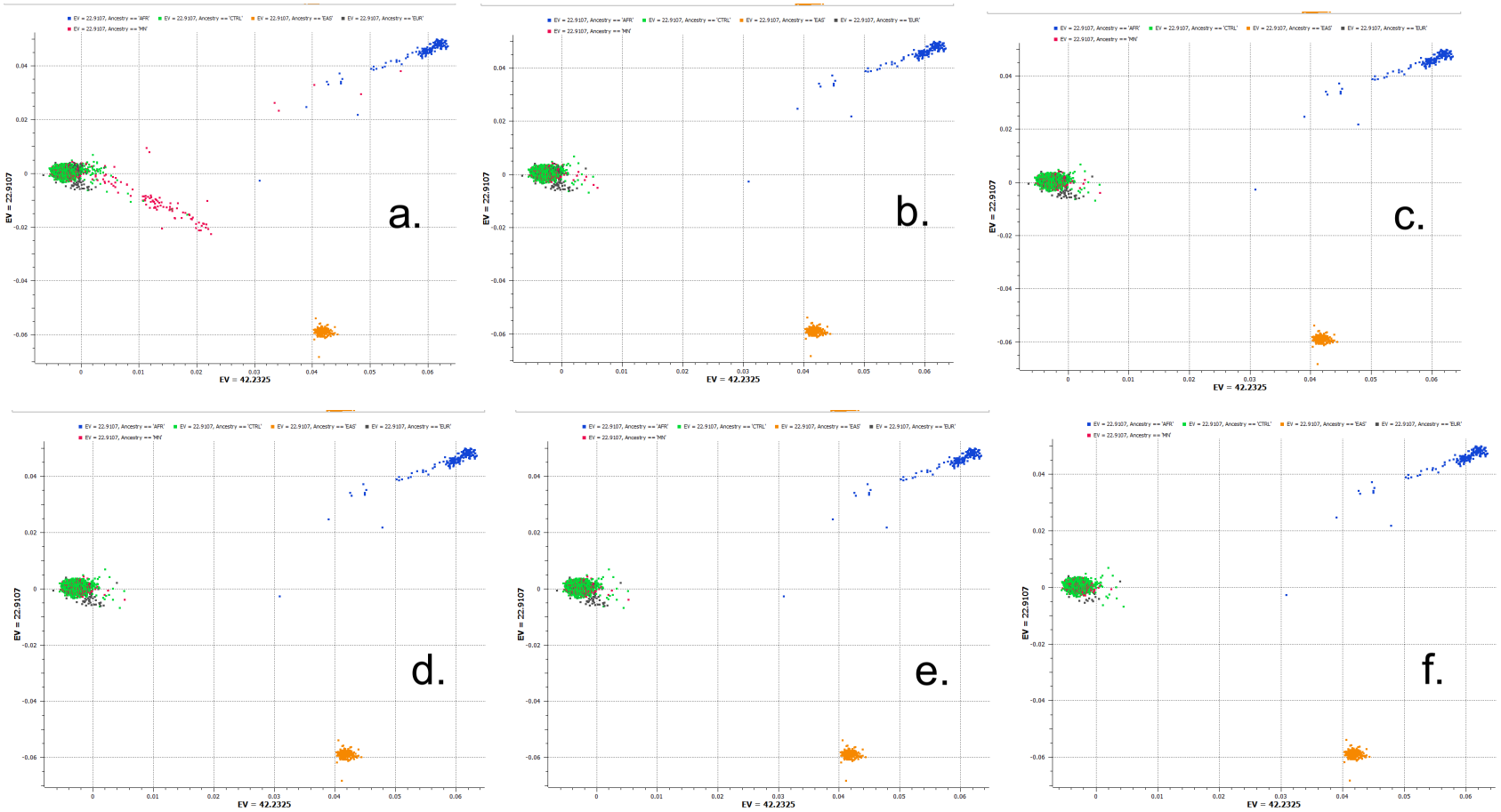


Figure 7.31: Principal component analysis of divergent ancestries in AMN cases and controls with the ancestry controls from 1000 Genomes Project. a. Unfiltered PCA b. Filtered for SD 4.0 c. Filtered for SD 3.5 d. Filtered for SD 3.25 e. Filtered for SD 3.0 f. Filtered for SD 2.5. AMN cases are highlighted in red, controls in green, European ancestry controls in black, African ancestry controls in blue and East Asian ancestry controls in orange.

7.2.3.2. Multidimensional scaling

MDS was done as an additional check for ancestry outlier exclusion. For MDS I used the 1000 Genomes Project ancestry controls. The same filtered case-control-ancestral control dataset was examined for MDS, Figure 7.32. This showed good overlap with the European ancestry controls and so this dataset was used for downstream analyses.

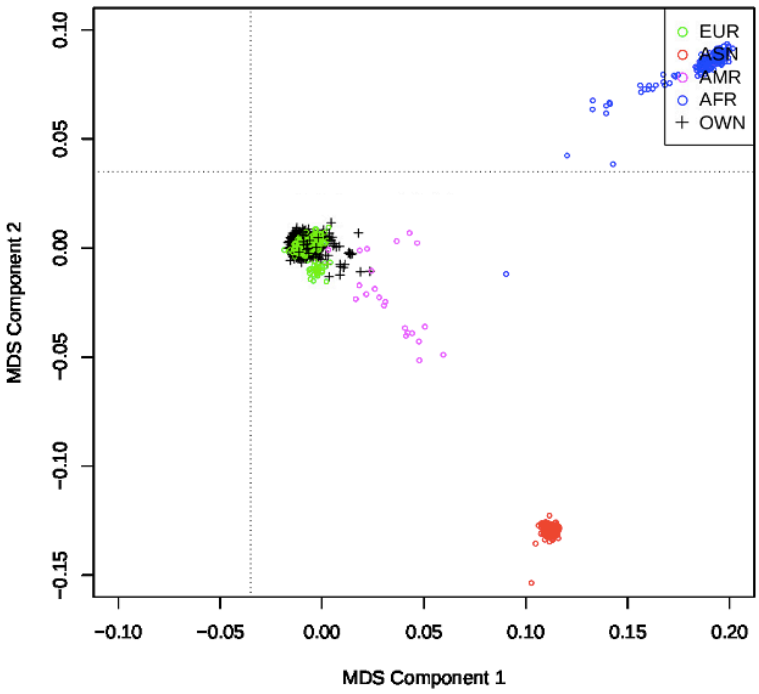


Figure 7.32: Multidimensional scaling analysis of divergent ancestries in AMN cases and reference controls with 1000 Genomes Project ancestry controls. AMN cases and controls as a combined dataset are shown in black crosses, European ancestry controls in green, African ancestry controls in blue, East Asian ancestry controls in red and admixed Americans in pink.

7.2.4. Imputation

7.2.4.1. Chromosome 2 imputation

Imputation was done with two different reference panels from the 1000 Genome Project phase 3 data; the 503 European or the 2504 all ancestry individuals. For

each of the different datasets I first imputed chromosome 2 only, the numbers of this analysis are summarised in Table 7.8.

	Pre-imputation	Conform-gt exclusion	Imputation reference panel	Post-imputation	Post-QC
AMN cases	28,788	20	European	2,384,141	316,997
			All ancestry	2,627,240	664,510
Controls	62,043	97	European	2,384,141	335,156
			All ancestry	2,409,325	518,843
Merged AMN-control	15,341	11	European	2,384,141	310,213
			All ancestry	2,409,325	167,116

Table 7.8: Overview of pre and post imputation of chromosome 2 with the different datasets and the different reference panels used.

7.2.4.1.1. Quality control post imputation

Post imputation QC is necessary to exclude poor quality imputed SNVs, see 6.2.5.1.2 and 6.2.6.4. The overview for each dataset remaining post QC is summarised in Table 7.8.

7.2.4.2. Whole genome imputation

Pre-imputation there were 188,662 QC filtered SNVs in the merged case-control dataset. In the European ancestry 503 individuals there are 27,520,389 SNVs in this reference panel across all chromosomes.

Post imputation all 22 chromosomes were merged and this resulted in 2,064,561 good quality SNVs across the whole genome in 1035 AMN cases and 5080 controls.

7.2.5. HLA imputation

HLA imputation was done with both the HapMap European reference panel (124 individuals, 3924 SNVs and 109 4-digit classical HLA alleles) and the T1DGC

reference panel (5225 individuals, 5868 SNVs and 298 4-digit classical HLA alleles), see 6.2.7.1. Since the T1DGC were first degree relatives of individuals with type 1 diabetes mellitus this might not be the best reference panel due to known HLA allele association in type 1 diabetes mellitus. This was the reason for the comparison of imputation with both reference panels.

With the T1DGC as the reference panel HLA imputation yielded 8962 SNV markers in the case and control dataset. With the HapMap reference panel the cases and controls were both imputed to 6701 SNV markers.

7.2.5.1. Quality control

For QC criteria see 6.2.7.4. R^2 score ≦0.8 excluded 171 SNV markers from the case and control dataset imputed with T1DGC reference panel and 2725 SNV markers from both case and control dataset imputed with the HapMap reference panel.

Imputation cannot accurately impute rare alleles, so MAF <1% was excluded. This excluded 815 SNV markers from the data imputed with the T1DGC reference panel and only a single SNV in those imputed with the HapMap reference panel.

Overall, there were 7976 SNV markers remaining post QC with the T1DGC reference panel and 3975 SNV markers with the HapMap as a reference panel. I wanted to assess only the 4-digit HLA types so excluded SNV markers that started with an rsID or the 1kg prefix. There were 115 4-digit HLA types with the T1DGC panel imputation and 56 with the HapMap reference panel remaining after QC, see Figure 7.33 for a summary.

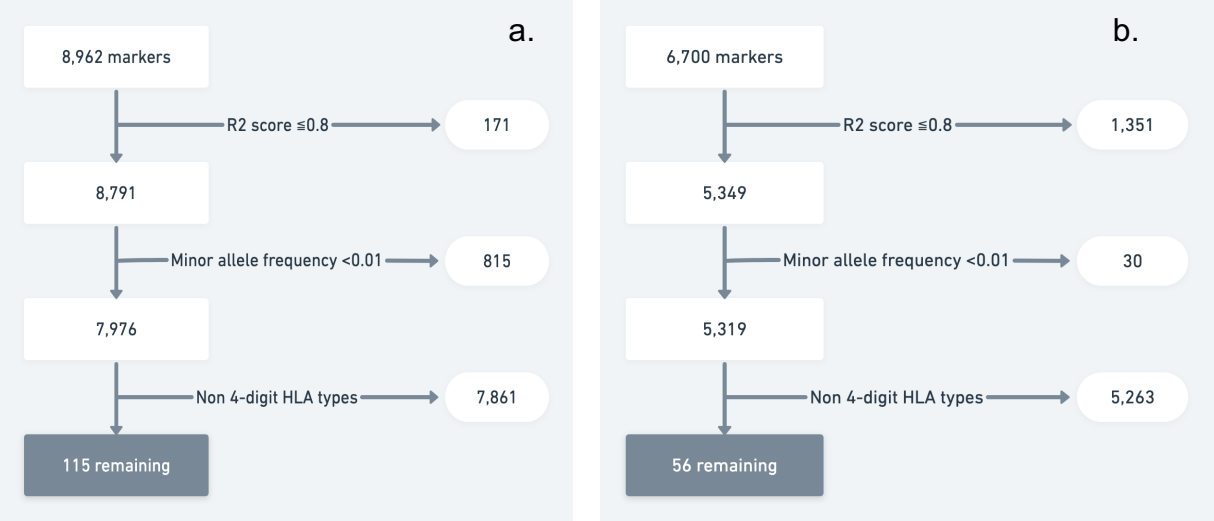


Figure 7.33: Flowchart summary of post HLA imputation quality control filtering. a. Filtering steps shown for the data imputed with the T1DGC reference panel. b. Filtering steps shown for the data imputed with the HapMap reference panel

7.2.6. Association tests

An additive model was used for association tests. The association test undertaken was logistic regression followed by logistic regression with additional co-variates with the top 10 principal components, to further account for population stratification and correct for minor differences within the case and control ancestry. Conditional analyses were undertaken on the lead identified SNVs.

7.2.6.1. Genome wide pre-imputation association test

To first explore the data prior to whole genome imputation I decided to run an association test with the post QC combined case-control dataset. This has a final number of 5080 controls and 1035 AMN cases. The number of intersecting SNVs was low at 188,662. An association test with logistic regression was done, Figure 7.34. With logistic regression the odds ratio for the lead SNV in chromosome 6, rs2854275 in *HLA-DQB1*, was 3.53 with a p-value of 4.19 $x10^{-95}$. The lead SNV in

chromosome 2, rs3792189 in *PLA2R1*, was 1.73 with a p-value of 2.37 $x10^{-28}$. Covariate analysis with the lead 10 principal components weakened the strength of the association but the two lead SNVs were still statistically significantly associated; *HLA-DQB1*, rs2854275 $p=1.6x10^{-24}$, OR 3.12 and *PLA2R1*, rs3792189 $p=2.96x10^{-17}$, OR 1.57.

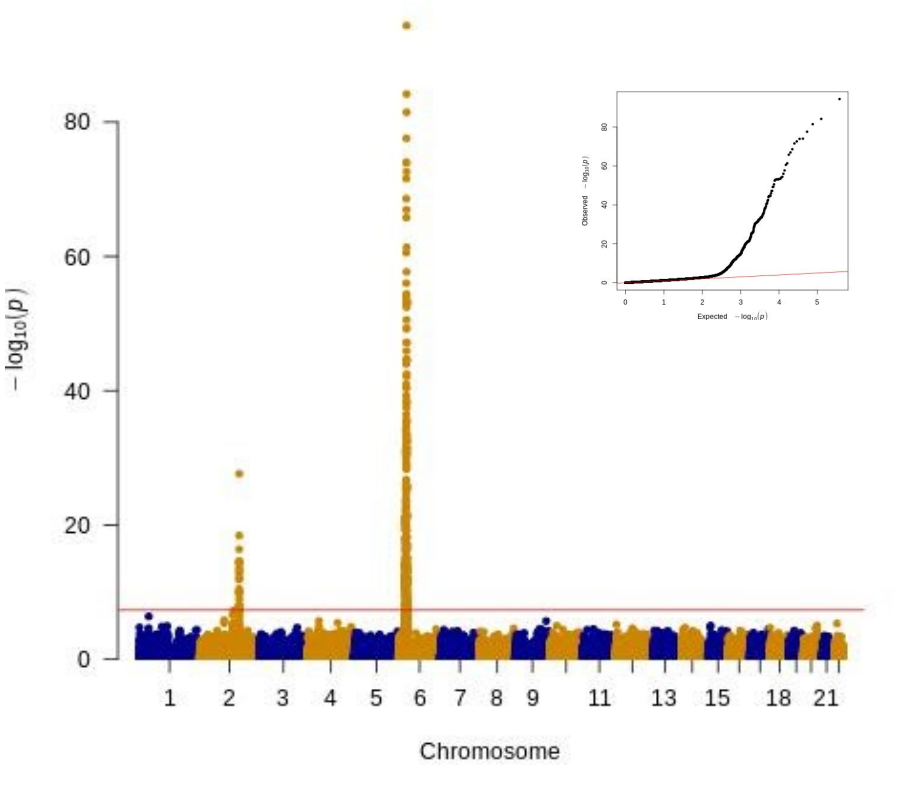


Figure 7.34: Manhattan plot of genome-wide logistic regression association test comparing European AMN cases to European controls. The x axis shows chromosome location, and the y axis shows negative decadic logarithmic p-values. Standardised genome-wide significance at $p=5x10^{-8}$ is depicted by the horizontal red line. The genomic inflation factor, lambda is 1.22. The inset graph is the QQ plot for observed vs expected p-values for this association test.

7.2.6.2. Chromosome 2 post-imputation association test

I undertook exploratory tests of the different imputation datasets in chromosome 2 alone. This was to save computational processing time prior to imputation across all the chromosomes. I wanted to ensure that the *PLA2R1* peak was replicable after

imputation and to ensure there was limited stratification which can result in a noisy Manhattan plot.

7.2.6.2.1. Datasets

After imputation of chromosome 2 there were 4 different datasets, summarised in Table 7.9.

Dataset number	Case and control	Imputation reference panel	Number of SNVs
1	Case and controls imputed separately	European	213,698
2	Case and controls imputed separately	All ancestry	353,361
3	Merged case and controls imputed together	European	167,116
4	Merged case and controls imputed together	All ancestry	310,213

Table 7.9: Overview of different post imputation datasets for chromosome 2.

7.2.6.2.2. Population stratification

Firstly, I calculated the genomic inflation factor to assess if there was potentially an issue caused by imputation that had affected population stratification. For dataset 1 the lambda had increased to 2.22, for dataset 2 =1.86, dataset 3 =1.23 (unchanged) and dataset 4 =1.27. It was at this stage due to these high genomic inflation factors I decided to undertake further principal component analysis to minimise this, see 7.2.3.1.2.1. After removal of 123 AMN cases and 107 controls I repeated the calculation of the genomic inflation factor; dataset 1 =1.68, dataset 2 =1.44, dataset 3 =1.25, dataset 4 =1.3. I didn't want to lose too many cases and controls and so decided to proceed with this because co-variate analysis with the lead principal components will further rectify the issue with population stratification.

7.2.6.2.3. Association analysis

A summary of the logistic regression association test is shown in Table 7.10:

Dataset	Lambda	Lead SNV	OR	p-value
1	1.68	*PLA2R1*: rs17341301	0.55	3.61×10^{-30}
2	1.44	*ZNF385B*: rs148306409	3.86	3.2×10^{-75}
3	1.25	*PLA2R1*: rs3792189	1.74	4.33×10^{-28}
4	1.3	Intergenic upstream of *PLA2R1*: rs4292050	0.55	1.45×10^{-28}

Table 7.10: Summary of the chromosome 2 logistic regression association test for the different datasets. OR = odds ratio.

These results demonstrated considerable deviation from the expected lead SNV and region from the pre-imputation association test apart from dataset 3. Further the odds ratio was inversed in dataset 1 and 4 whereas these are expected to be risk SNVs. As an additional check I reviewed the *.fam file to ensure that the phenotype had been coded correctly as 1 for controls and 2 for cases; this was correct and there was no difference when I specified the phenotype *.fam file from the dataset 3.

The high genomic inflation factor could have been contributory to the unexpected results so logistic regression with the 10 lead principal components as co-variates was done. The results are summarised in Table 7.11.

Dataset	Lead SNV	OR	p-value
1	Intergenic upstream of *SLC39A10*: rs34055576	0.55	3.61×10^{-30}
2	*ZNF385B*: rs148306409	3.95	3.28×10^{-76}
3	*PLA2R1*: rs3792189	1.74	1.47×10^{-27}
4	Intergenic upstream of *PLA2R1*: rs4292050	0.55	2.27×10^{-27}

Table 7.11: Summary of the chromosome 2 logistic regression association test with the 10 lead principal components as an additional covariate for the different datasets. OR = odds ratio.

The results had not changed significantly or improved despite the additional co-variate with principal components.

A final idea I had was datasets 2 and 4 had been imputed with the all ancestry reference panel which is known to be less accurate for rare SNVs. So, I decided to perform further MAF filtering and exclude any SNVs with a MAF <5% to further minimise any imputation error. To enable direct comparison, I decided to also perform the MAF filtering on the datasets imputed with a European reference panel (dataset 1 and 3). Further logistic regression association tests were done which did not change the results for datasets 1, 3 and 4. For dataset 2, the lead SNV did change to another gene which was not near *PLA2R1*. Interestingly in dataset 2 the second lead SNV was in *PLA2R1*, rs4665138, with an odds ratio of 0.54 and a p-value =8.03 $x10^{-34}$. The results for the lead SNVs with a MAF >5% are summarised in Table 7.12:

Dataset	Lead SNV	OR	p-value
1	*PLA2R1*: rs17341301	0.55	3.62 $x10^{-30}$
2	*DNAJC27-AS1*: rs2196792	2.02	1.55 $x10^{-42}$
3	*PLA2R1*: rs3792189	1.74	4.33 $x10^{-28}$
4	Intergenic upstream of *PLA2R1*: rs4292050	0.55	1.45 $x10^{-28}$

Table 7.12: Summary of the chromosome 2 logistic regression association test only including SNVs with a minor allele frequency >5%. OR = odds ratio.

After direct comparison of all the results and the differences between the expected region of interest from the pre-imputed data and the previously reported lead chromosome 2 SNV in Stanescu *et al*. [73] I decided that the most reliable dataset to proceed for whole genome imputation was dataset 3. This is the merged AMN and control dataset imputed with the European reference panel. The results were consistent despite the different remedial measures trialled and it also matches the results from the pre-imputation association analysis. From this point on I chose to exclusively analyse dataset 3.

7.2.6.3. Whole genome post-imputation association test

7.2.6.3.1. Logistic regression

Logistic regression was done with the 1035 AMN cases compared to 5080 controls of European ancestry. After imputation there were 2,064,561 SNVs spanning the whole genome for association analyses. There were two significantly associated regions; one in chromosome 2 and another in chromosome 6, Figure 7.35. The lead SNV in the chromosome 2 peak was in *PLA2R1*, rs3792189, OR 1.72, p-value =3.5 $x10^{-28}$, with a A>C substitution. The lead SNV in chromosome 6 was in *HLA-DQA1*, rs9272532, OR 3.29, p-value =1.16 $x10^{-102}$, with a C>A substitution. Other regions not reaching statistical significance are; chromosome 1, rs6675787 in *HPCAL4*, OR 1.29, p-value =4.52 $x10^{-7}$; chromosome 4, rs9918077 intergenic region and closest genes are *TRPC3* and *KIAA1109*, OR 0.69, p-value =8.16$x10^{-7}$; chromosome 9, rs670028 in *DENND1A*, OR 1.39, p-value=4.0 $x10^{-7}$.

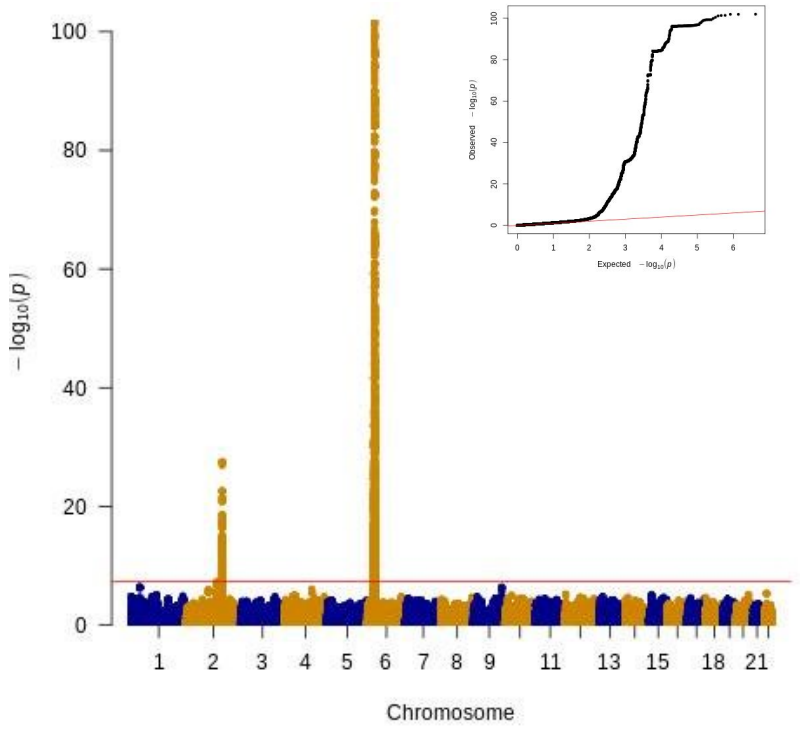


Figure 7.35: Manhattan plot of genome-wide logistic regression association test comparing 1035 European AMN cases to 5080 controls.
The x axis shows chromosome location, and the y axis shows negative decadic logarithmic p-values. Standardised genome-wide significance at $p=5x10^{-8}$ is depicted by the horizontal red line. The genomic inflation factor, lambda is 1.24. The inset graph is the QQ plot for observed vs expected p-values for this association test.

7.2.6.3.2. Principal component co-variate analysis

The top 10 principal components were calculated and utilised as an additional co-variate in the logistic regression association analysis. This did not change the results of the association test and the same two lead SNVs were still associated with AMN. In *PLA2R1* on chromosome 2, rs3792189, the odds ratio was slightly lower at 1.63 and the p-value had increased to 1.95×10^{-21} which was still statistically significant. The lead SNV in *HLA-DQA1* on chromosome 6 also remained, rs9272532, OR 2.6 and p-value increased considerably to 1.29×10^{-29}.

7.2.6.3.3. Conditional analyses

Conditional analyses on the lead SNVs within the identified GWAS peaks were conducted in a stepwise manner to identify any further independent SNVs within either the chromosome 2 or chromosome 6 peaks as seen in the Manhattan plot. The first conditional analysis was done on the lead SNV in chromosome 6, rs9272532. The chromosome 2 lead SNV, rs3792189, in *PLA2R1* was the second independently associated SNV with an odds ratio of 1.68 and p-value =1.28 $x10^{-23}$, see Figure 7.36

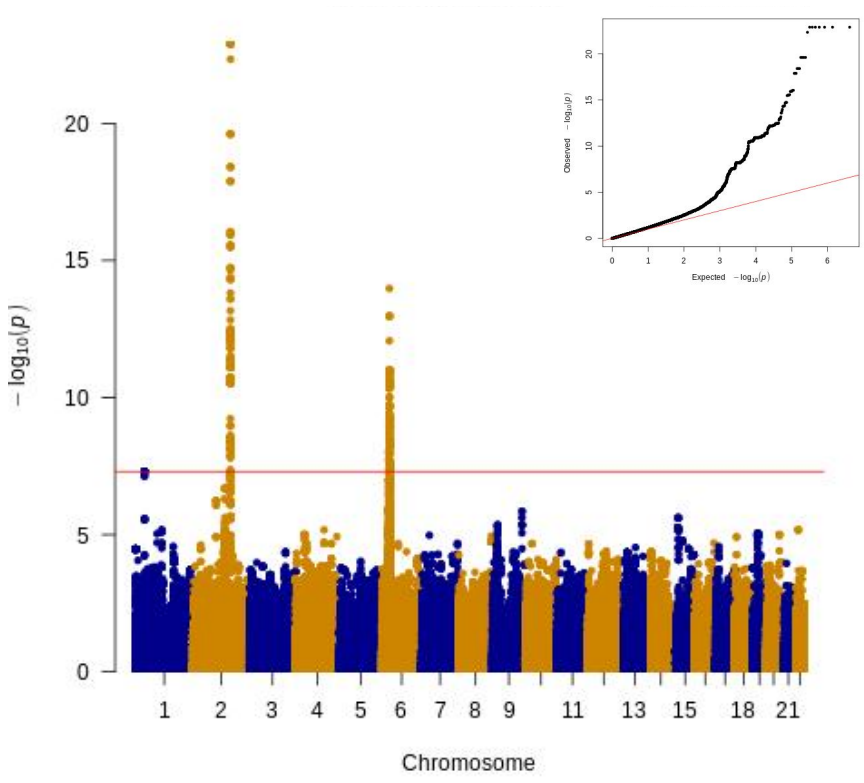


Figure 7.36: Manhattan plot of genome-wide logistic regression association test conditioned on the lead SNV, rs9272532. Comparison of 1035 European AMN cases to 5080 controls. The x axis shows chromosome location, and the y axis shows negative decadic logarithmic p-values. Standardised genome-wide significance at $p=5x10^{-8}$ is depicted by the horizontal red line. The inset graph is the QQ plot for observed vs expected p-values for this association test.

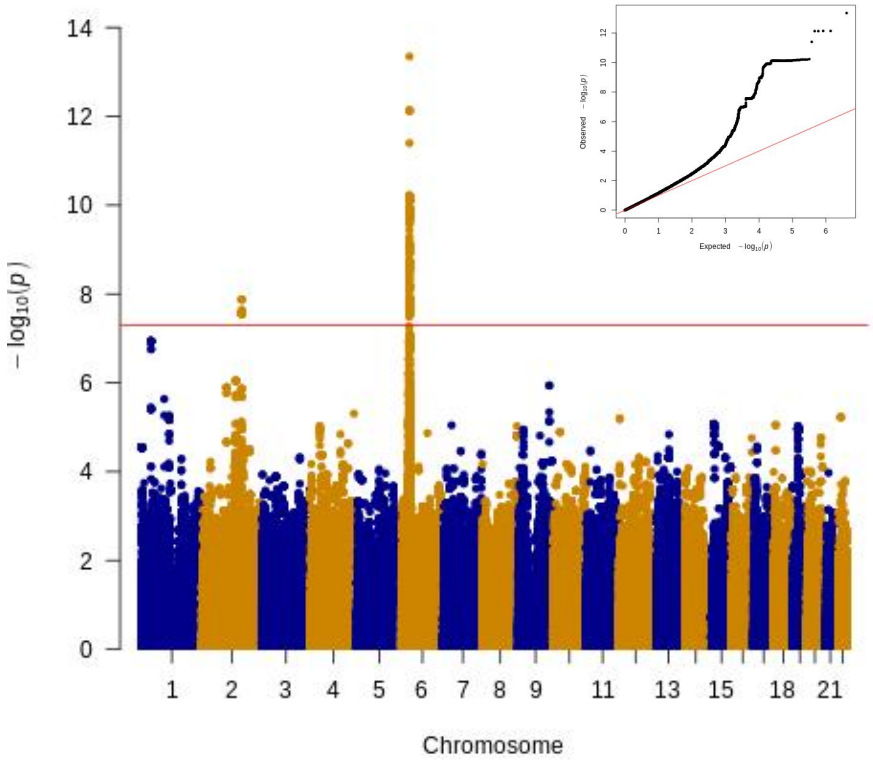


Figure 7.37: Manhattan plot of genome-wide logistic regression association test conditioned on the lead 2 SNVs, rs9272532 and rs3792189. Comparison of 1035 European AMN cases to 5080 controls.
The x axis shows chromosome location, and the y axis shows negative decadic logarithmic p-values. Standardised genome-wide significance at $p=5x10^{-8}$ is depicted by the horizontal red line. The inset graph is the QQ plot for observed vs expected p-values for this association test.

Conditional analysis done on the lead 2 SNVs in both chromosome 2 and 6 identified a second SNV in the chromosome 6 peak; rs9469220 which is intergenic but the closest gene is *HLA-DQB1*. The odds ratio is 0.66 suggesting it is a protective SNV, p-value =4.46 $x10^{-14}$, see Figure 7.37.

Conditional analysis done on the lead 3 SNVs in both chromosome 2 and 6 identified a second SNV in the chromosome 2 peak; rs2292390 in *PLA2R1*, OR 1.57, p-value =1.31 $x10^{-8}$, see Figure 7.38a. Conditional analysis done on the 4 independent SNVs

resulted in no further statistically significant associations. The remaining non significantly associated lead SNV was in chromosome 1, rs538758, in *HPCAL4*, OR 1.34, p-value =8.29 x10^{-8}, see Figure 7.38b.

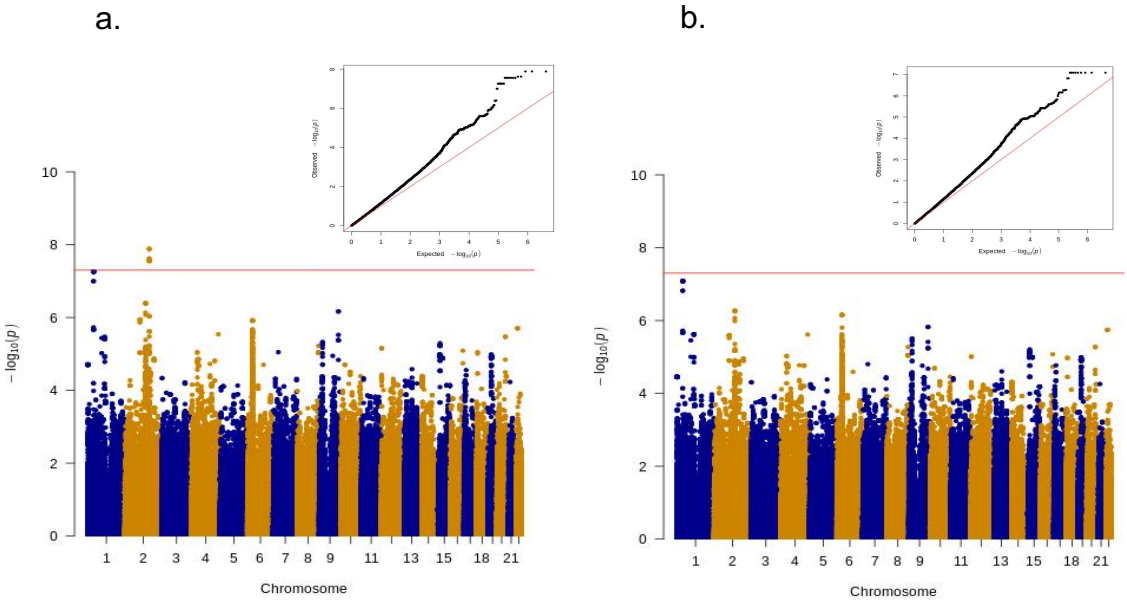


Figure 7.38: Manhattan plot of genome-wide logistic regression conditional association tests. 1035 AMN cases compared against 5080 controls. Inset graphs are the associated QQ plot for that test. Standardised genome-wide significance at p=5x10^{-8} is depicted by the horizontal red line.

a. Conditioned on the lead 3 SNVs, rs9272532, rs3792189 and rs9469220.
b. Conditioned on lead 4 SNVS, rs9272532, rs3792189, rs9469220 and rs2292390

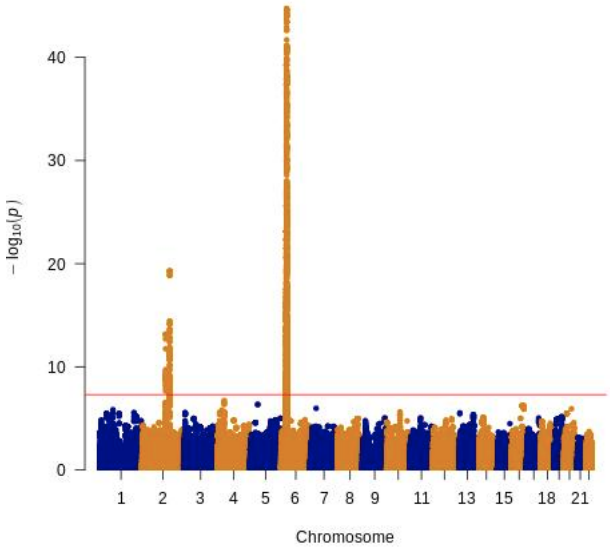
7.2.6.4. Antibody status

With the phenotype data I examined genome wide association for the specific sub-antibody groups.

7.2.6.4.1. aPLA2Rab positive versus controls

Extracting aPLA2Rab AMN cases and healthy controls resulted in 372 cases and 4929 controls. The logistic regression association test demonstrated an association in two identified regions; *HLA-DQA1* and *PLA2R1*, Figure 7.39, with both SNVs being the same as in 7.2.6.3.1. The lead SNV in *HLA-DQA1* is rs9272532, OR =4.78 and p =2.83x10^{-74}. The lead SNV in chromosome 2 is rs3792189, OR =2.03 and p= 5.42x10^{-20}. There is a demonstrable peak visible in chromosome 4, the lead SNV here is rs73236604 which is close to *TLR10* but it does not reach statistical significance, p=1.94x10^{-7} and OR 1.79.

Figure 7.39: Manhattan plot of genome-wide logistic regression association tests. 372 aPLA2Rab AMN cases compared against 4929 controls. Standardised genome-wide significance at p=5x10^{-8} is depicted by the horizontal red line.



7.2.6.4.2. Anti-THSD7A antibody positive versus controls

The anti-THSD7A antibody AMN cases and healthy controls were merged; resulting in 31 AMN cases and 4929 controls. A logistic regression for association did not demonstrate any statistically significant associations, Figure 7.40.

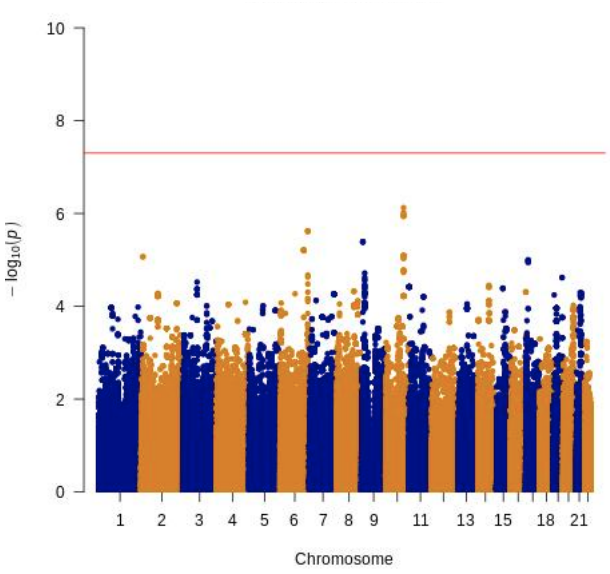


Figure 7.40: Manhattan plot of genome-wide logistic regression in 31 anti-THSD7A AMN cases compared against 4929 controls. Standardised genome-wide significance at p=5x10^{-8} is depicted by the horizontal red line.

7.2.6.4.3. Anti-THSD7A antibody positive versus aPLA2Rab

This analysis compared 31 anti-THSD7A antibody positive AMN cases against 372 aPLA2Rab positive AMN cases. There were no statistically significant associations and the lead SNV was rs9272343 which is in *HLA-DQA1*, p=1.07x10^{-6}.

7.2.6.4.4. Dual negative antibody versus controls

Comparing 355 dual negative antibody cases against 4929 controls demonstrated a statistically significant association in both *HLA-DQA1* and *PLA2R1*. The lead association in chromosome 6 is in rs6932167 with an OR of 2.86, p=2.61x10^{-29} and in chromosome 2 is rs3792189, OR of 2.86, p=1.58x10^{-9} which is the same lead SNV as the previously identified *PLA2R1* SNV in this study, 7.2.6.3.1, Figure 7.41. See 8.2.4.5 for specific limitations.

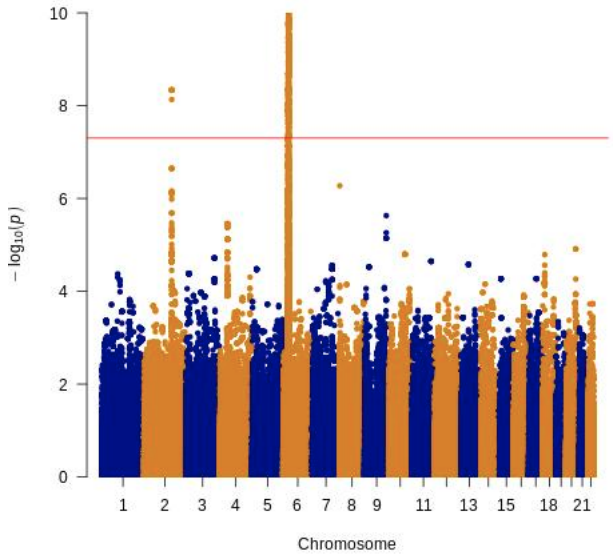


Figure 7.41: Manhattan plot of genome-wide logistic regression in 355 dual antibody negative AMN cases compared against 4929 controls. Standardised genome-wide significance at $p=5x10^{-8}$ is depicted by the horizontal red line.

7.2.6.5. Clinical parameters

Relevant important clinical phenotype information was available in a small subset of 243 aPLA2Rab positive AMN individuals from our German cohort, of which 225 were within the post QC dataset. I decided to investigate if there was any genetic association with these clinical parameters at a genome wide level in the imputed post QC dataset. Because all of these individuals had been genotyped on the same microarray I was able to use all the post imputation post QC SNV markers for these tests; 2,064,561 SNVs in the German cohort and 5,132,131 SNVs in the historical 2011 AMN cohort.

7.2.6.5.1. Association with glomerular filtration rate

Estimated glomerular filtration rate decline is a useful analysis as treatment for AMN involves in preserving kidney function. Data was available in 225 European AMN individuals for eGFR across different time points. I calculated the average eGFR decline per year in individuals only with data for more than 5 years. The eGFR decline was examined with a linear regression association test. There was no association between eGFR decline (mL/min/1.73m^2/year) in the German cohort, Figure 7.42.

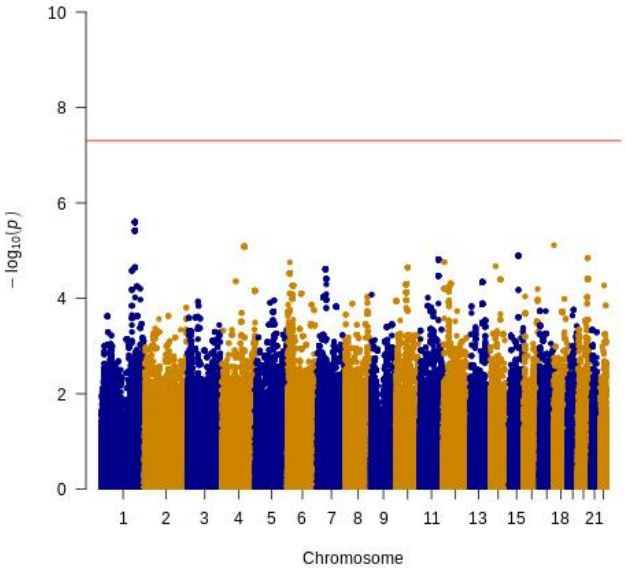


Figure 7.42: Manhattan plot of genome-wide association test with linear regression with average eGFR decline / year (mL/min/1.73m^2/year) Comparison of 225 German, European AMN cases. The x axis shows chromosome location, and the y axis shows negative decadic logarithmic p-values. Standardised genome-wide significance at p=5x10^{-8} is depicted by the horizontal red line.

7.2.6.5.2. Association with other correlates of poor renal outcomes

The main interest was to see if there were any other good genetic predictors for individuals that have poor renal outcomes. If so, this would enable a tool to identify individuals that may require early treatment to preserve their renal function.

Association tests with other known predictors of poor renal outcomes (at diagnosis); uPCR (mg/mmol), anti-PLA2R antibody titres (kunits/L), eGFR at presentation (mL/min/1.73m^2), age and gender were undertaken. Immunosuppression data was available and this was added as a co-variate to account for individuals that had treatment and therefore had an intervention to protect renal function. These association tests are demonstrated in Figure 7.43. Statistically significant associations were found with uPCR on chromosome 11 (rs637148, p-value 1.89 x10^{-8}) and with aPLA2Rab on chromosome 4 (rs6840215, p-value 3.99 x10^{-8}). No other statistically significant associations were found with the other clinical parameters.

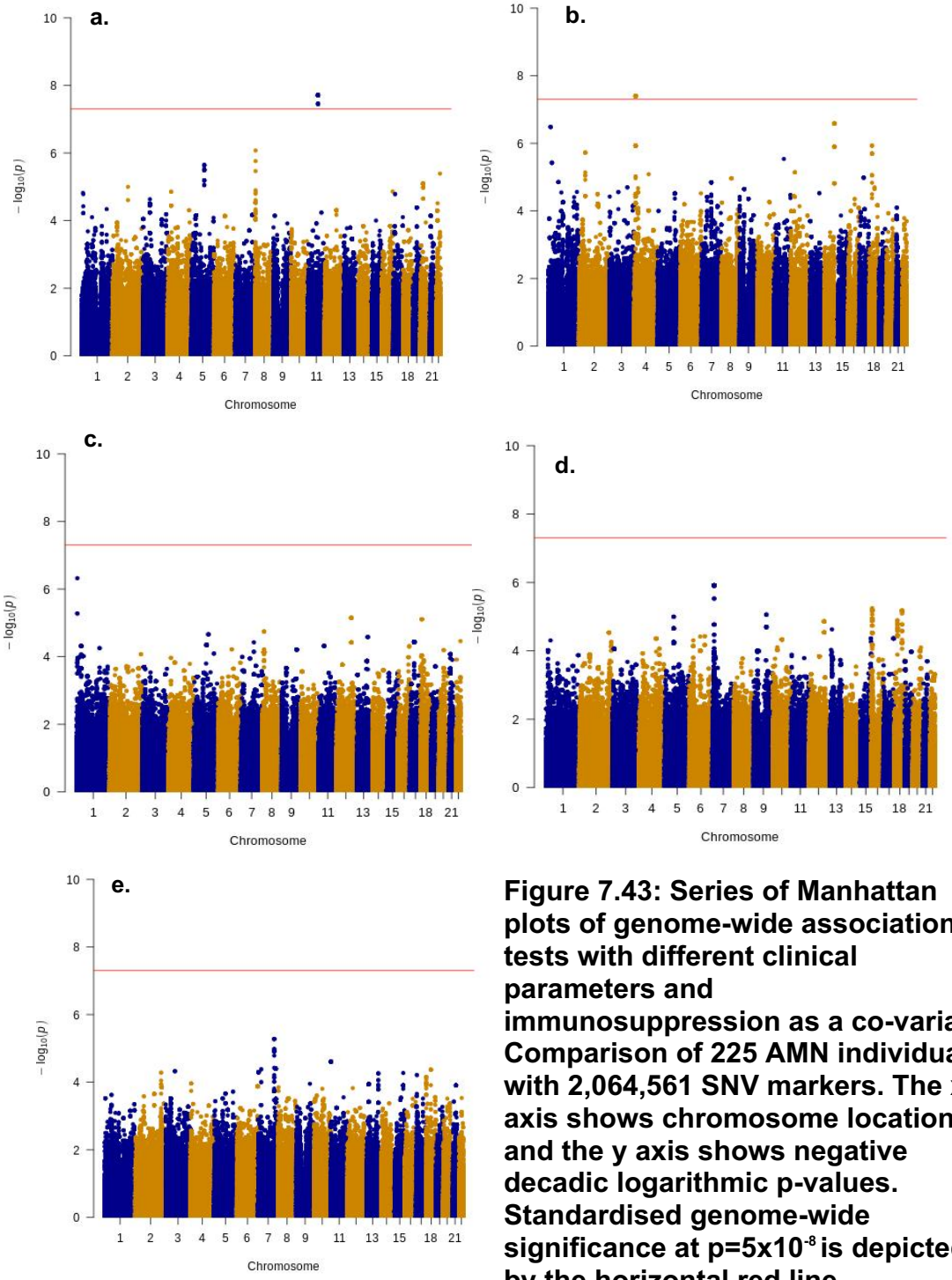


Figure 7.43: Series of Manhattan plots of genome-wide association tests with different clinical parameters and immunosuppression as a co-variate. Comparison of 225 AMN individuals with 2,064,561 SNV markers. The x axis shows chromosome location, and the y axis shows negative decadic logarithmic p-values. Standardised genome-wide significance at $p=5x10^{-8}$ is depicted by the horizontal red line.
a. Linear regression of baseline uPCR (mg/mmol)
b. Linear regression of anti-PLA2R antibody titres (kunits/L)
c. Linear regression of baseline eGFR (mL/min/1.73m^2)
d. Linear regression of age
e. Logistic regression of gender

7.2.6.6. HLA association test

The HLA association test was performed with only the 4-digit HLA types. Owing to the issues described in 6.2.7.1 I undertook the association tests in both imputed datasets (T1DGC reference panel and HapMap reference panel). With the T1DGC reference panel in the case control dataset there were 115 4-digit HLA types, and with the HapMap reference panel there were 56. Two main analyses for HLA association were done; firstly all aPLA2Rab positive AMN cases against controls, secondly all anti-THSD7A antibody AMN cases against controls.

7.2.6.6.1. Anti-PLA2R antibody AMN versus controls

The HLA association test for anti-PLA2R antibody AMN had 372 cases compared against 4717 controls.

7.2.6.6.1.1. HLA imputation with the T1DGC

The strongest signal was observed in HLA-DQA1*05:01 (p-value =$1.02x10^{-68}$, odds ratio 4.38), Table 7.13. The second lead signal was in HLA-DRB1*03:01 (p-value =$1.67x10^{-64}$, odds ratio 4.73). These are the two same HLA types identified in the most recent GWAS by Xie *et al.* [171]. To overcome population stratification, I undertook the association test with correction for the 10 lead principal components. This still identified the same two lead HLA types, but the strength of the association and odds ratio reduced, Table 7.14. Finally, a conditional analysis with HLA-DQA1*0501 and the 10 lead principal component analysis confirmed that the association with HLA-DRB1*03:01 was independent with an OR =2.64 and p-value =2.31 $x10^{-7}$. This should be interpreted with care as discussed in 5.2.8 as these two HLA types are in tight linkage and so can not be accurately delineated.

Unconditioned HLA analysis					
HLA type	**OR**	**p-value**	**CI (95%)**	**AF controls**	**AF cases**
HLA-DQA1*05:01	4.39	1.02E-68	3.72 - 5.17	0.25	0.55
HLA-DRB1*03:01	4.73	1.67E-64	3.95 - 5.66	0.15	0.39
HLA-DQB1*02:01	4.6	6.65E-63	3.85 - 5.5	0.15	0.39
HLA-B*08:01	3.57	7.25E-42	2.97 - 4.3	0.15	0.32

Table 7.13: HLA association test for anti-PLA2R antibody AMN against controls, data imputed with the T1DGC reference panel. This is an unconditioned logistic regression association analysis. Results shown are HLA type, odds ratio (OR), p-value, confidence interval (CI) and allele frequencies (AF). Data are sorted by p-values.

HLA association corrected for 10 principal components			
HLA type	**OR**	**p-value**	**CI (95%)**
HLA-DQA1*05:01	3.21	2.35E-20	2.51 - 4.11
HLA-DRB1*03:01	4.31	1.04E-15	3.02 - 6.17
HLA-DQB1*02:01	3.83	4.25E-14	1.4 - 3.33
HLA-DQA1*03:01	0.35	5.98E-08	1.0 - 3.33

Table 7.14: HLA association test for anti-PLA2R antibody AMN against controls with correction for the 10 lead principal components, data imputed with the T1DGC reference panel. Results shown are the HLA type, odds ratio (OR), p-value, confidence interval (CI). Data are sorted by p-values.

7.2.6.6.1.2. HLA imputation with the HapMap reference panel

With the data imputed with the HapMap reference panel the strongest signal was observed in HLA-C*07:01, p-value =4.81x10^{-46}, OR =3.46. The second strongest association was in HLA-DRB1*03:01, p-value 1.35x10^{-41}, OR =3.41. The results of the lead associations are shown in Table 7.15. With correction for the 10 lead principal components the associations were considerably different with HLA-DQB1*05:03 being the lead variant, p-value =3.24x10^{-10}, OR =3.14. The second strongest association was in HLA-A*03:01, p-value 2.24x10^{-6}, OR =2.73,

summarised in Table 7.15. These are different to those previously reported and different to the data imputed with the T1DGC reference panel.

Unconditioned HLA analysis			**HLA association corrected for 10 principal components**		
HLA type	**OR**	**p-value**	**HLA type**	**OR**	**p-value**
HLA-C*07:01	3.46	4.81E-46	HLA-DQB1*05:03	3.14	3.24E-10
HLA-DRB1*03:01	3.42	1.35E-41	HLA-A*03:01	2.73	2.24E-6
HLA-DQA1*05:01	2.96	1.76E-41	HLA-C*07:01	1.93	3.97E-6
HLA-B*08:01	3.52	4.96E-41	HLA-DRB1*15:01	2.52	5.64E-6

Table 7.15: HLA association test for anti-PLA2R antibody AMN against controls, data imputed with the HapMap reference panel. Unconditioned data analysis is shown on the left. On the right the conditioned data is conditioned only for the top 10 lead principal components. Results shown are the HLA type, odds ratio (OR) and the p-value. Data are sorted by p-values.

7.2.6.6.2. Anti-THSD7A antibody AMN versus controls

I was interested in determining the HLA type for the anti-THSD7A antibody AMN cases and so explored this with an HLA association test with the dataset comprising of 32 cases compared against 4717 controls.

7.2.6.6.2.1. HLA imputation with the T1DGC

An HLA association was found with anti-THSD7A antibody AMN cases and HLA-DRB1*08:01 in the unconditioned analysis. After correction for the 10 lead principal components none of the HLA types reached statistical significance (p-value <0.00043). Data for the lead HLA types are summarised in Table 7.16.

Unconditioned HLA analysis			HLA association corrected for 10 principal components		
HLA type	**OR**	**p-value**	**HLA type**	**OR**	**p-value**
HLA-DRB1*08:01	79.8	8.53E-5	HLA-DPB1*13:01	14	0.0037
HLA-DQB1*02:02	2.74	0.0005			

Table 7.16: HLA association test for anti-THSD7A antibody AMN against controls, data imputed with the T1DGC reference panel. Unconditioned data analysis is shown on the left. On the right the conditioned data is conditioned only for the top 10 lead principal components. Results shown are the HLA type, odds ratio (OR) and the p-value. Data are sorted by p-values.

7.2.6.6.2.2. HLA imputation with the HapMap reference panel

With the data imputed with the HapMap reference panel the strongest signal was in a class I HLA type; HLA-B*44:03, p-value =0.0011, OR 29.46. With principal component correction the lead HLA type was HLA-DRB1*11:01, p-value =0.01 and OR =1.14. Neither of these were statistically significant.

7.2.7. Epistasis

It was important to examine the historical 2011 GWAS data so that the genetic risk score could be calculated correctly as there should be no epistasis in an additive model. Utilising the lead 2 SNVs from the 2011 AMN study a test for epistasis in PLINK v1.9 did not demonstrate any interaction between the two loci. The odds ratio for interaction between the 2 SNVs (rs4664308 and rs2187668) was 0.88 and the p-value =0.35.

I also analysed my current data for epistasis between the four lead SNVs identified in this GWAS. To facilitate and reduce computational processing time I undertook pairwise interaction tests between each of these 4 SNVs only. In PLINK this demonstrated epistasis between the 2 lead SNVs in chromosome 2 and chromosome 6 in the combined case and control dataset, p-value =9.25 $x10^{-5}$

(statistically significant p-value =0.0083 (0.06/6)) with an odds ratio for interaction of 1.38. Assessing for epistasis in AMN cases only also confirmed epistasis was present with a p-value =2.64 $x10^{-5}$.

To confirm there was epistasis between the two lead SNVs an alternative method for epistasis was undertaken. Because PLINK had already identified interaction between the 2 lead SNVs I decided to check for interaction in these 2 SNVs only. Utilising the W-test this also demonstrated epistasis between the 2 lead SNVs, rs9272532 and rs3792189. The odds ratio is 1.38 for interaction and the p-value =1.43 $x10^{-127}$. The W-test statistics show a good estimation of the parameters at each of the 9 different combinations of genotypes, see Figure 7.44.

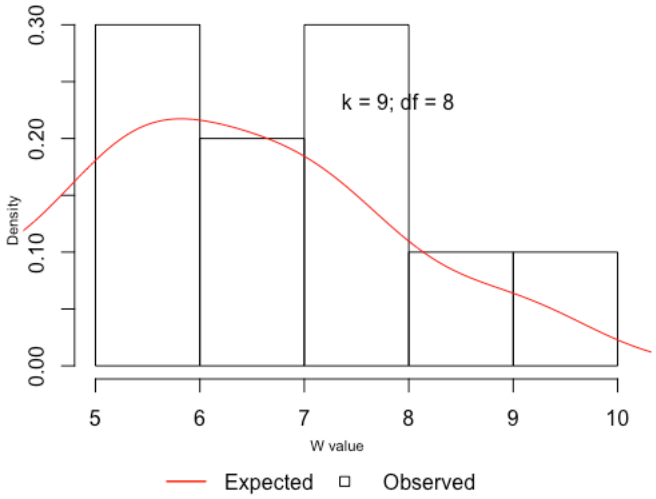


Figure 7.44: The expected versus observed histogram of the W-test statistics at each combination size (k) with degrees of freedom (df).

7.3. Genetic risk score

The GRS was calculated from the previously identified 2011 independent risk variants associated with AMN. My analysis utilised the odds ratio from this previous GWAS. In the 2011 data there was no epistasis and for this reason I used an additive model for the genetic risk. For each individual from the genotyped case dataset the GRS was calculated, see 6.3 for further details.

7.3.1. Case-control dataset

There were 1102 European MN cases; 823 had known status for both anti-PLA2R and anti-THSD7A antibodies. The composition of the cohort is detailed in Table 7.17. Through collaborators, I also had access to DNA from a small and very rare AMN cohort of 15 non-familial European paediatric cases of dual antibody negative AMN as well as 7 anti-contactin antibody associated MN cases; these two groups are examined separately.

Healthy Controls	**5642**
Steroid sensitive nephrotic syndrome	**422**
MN: Anti-PLA2R1 antibody positive	406
MN: Anti-THSD7A antibody positive	32
MN: Dual antibody (PLA2R1 & THSD7A) positive	1
MN: Dual antibody (PLA2R1 & THSD7A) negative	384
MN: Serum not available	279
Total Membranous nephropathy	**1102**

Table 7.17: Composition of cohort (AMN, membranous nephropathy; PLA2R1, phospholipase A2 receptor-1; THSD7A, thrombospondin type-7 domain containing 1A)

7.3.2. Antibody group and genetic risk

Compared with the GRS in 5642 healthy controls (median =0.17) and 422 SSNS controls (median =0.17), GRS was significantly elevated in the aPLA2Rab group (N =406; median =0.67; p <0.0001), Figure 7.45. GRS in the 384 individuals with dual antibody negative MN and the 279 individuals in whom serum was unavailable was 0.34; intermediate and statistically significantly different from, the control (p <0.0001) and PLA2R1-positive groups (p <0.0001), consistent with them comprising a mixture of PLA2R1-positive and negative individuals. The median GRS among the 406 with aPLA2R1ab was significantly higher than among the 32 with THSD7A antibodies (0.17, p <0.0001).

7.3.3. Allele count

The risk allele counts at both the *PLA2R1* and *HLA-DQA1* loci were compared to determine their individual contribution to the genetic risk of AMN.

7.3.3.1. *PLA2R1* risk allele

Corroborating previous studies [91], the data demonstrated that the *PLA2R1* risk allele number was significantly different between healthy subjects and the aPLA2Rab group (p <0.0001), Figure 7.46. Healthy subjects were also significantly different from the dual antibody negative (p <0.0001) and the serum unavailable (i.e. antibody status unknown) groups (p <0.0001). The SSNS group had the similar risk allele frequency as healthy controls and were statistically significantly different to the same three groups: aPLA2Rab group, dual antibody negative and serum unavailable (p <0.0001 in all groups). The dual antibody negative and serum unavailable groups

are a mixed group of individuals and are statistically different compared to the aPLA2Rab group (p =0.0007 and p =0.0003, respectively). When correcting for the multiple comparisons in this study, the difference in *PLA2R1* risk allele number was not statistically significant (p =0.04) between anti-THSD7A antibody cases and aPLA2Rab cases.

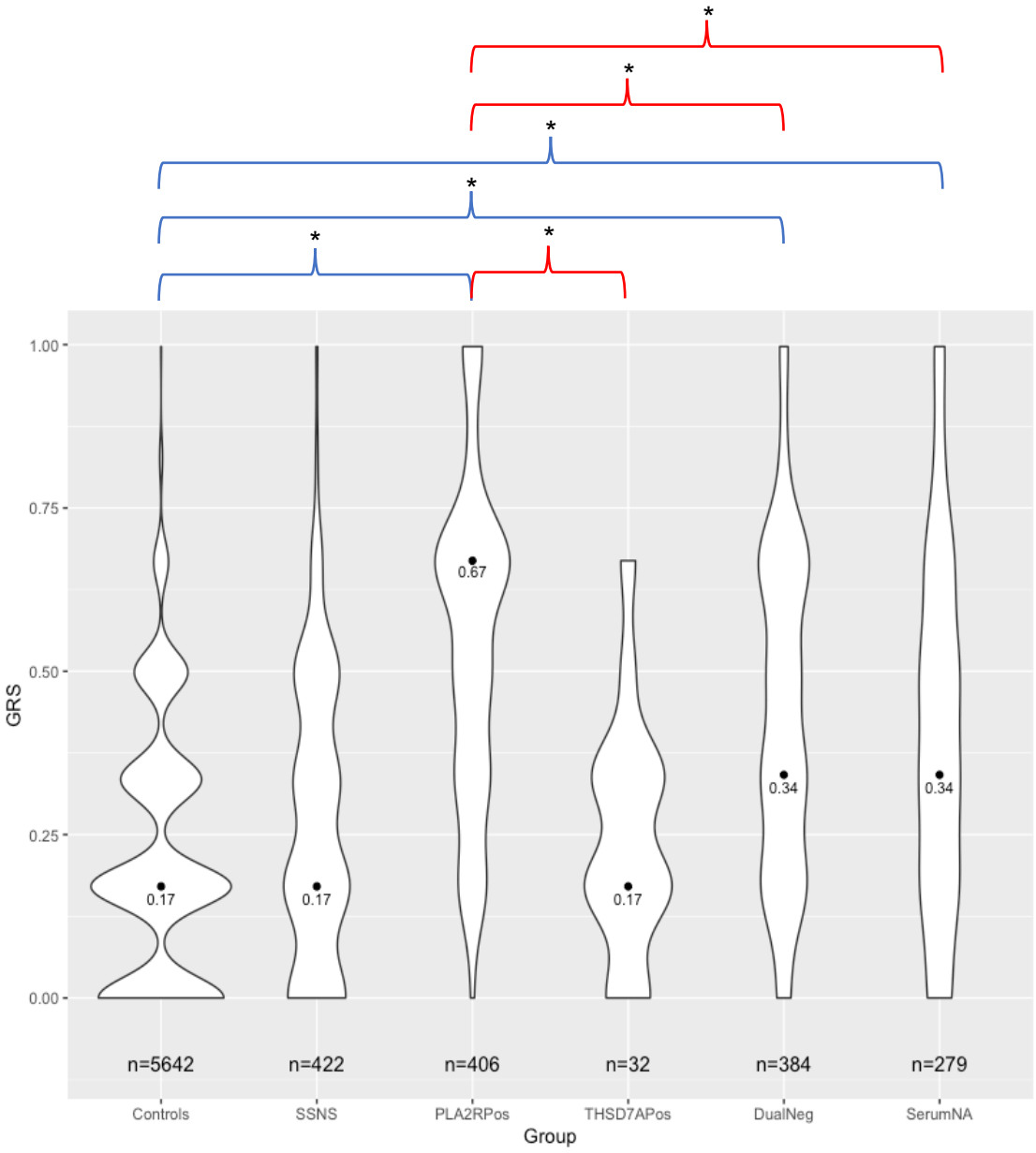


Figure 7.45: Violin plot of genetic risk score in different phenotype groups. Number of individuals in each group is shown. The black dot represents the median. Statistically significant differences from Controls are highlighted by blue braces, (p <0.0001). Statistically significant differences from PLA2Rpos are highlighted by red braces, (p <0.0001). SSNS, steroid sensitive nephrotic syndrome; PLA2RPos, anti-PLA2R1 antibody positive AMN; THSD7APos, anti-THSD7A antibody positive AMN; DualPos, PLA2R1 and THSD7A antibody positive MN; DualNeg, PLA2R1 and THSD7A antibody negative AMN; SerumNA, serum not available.

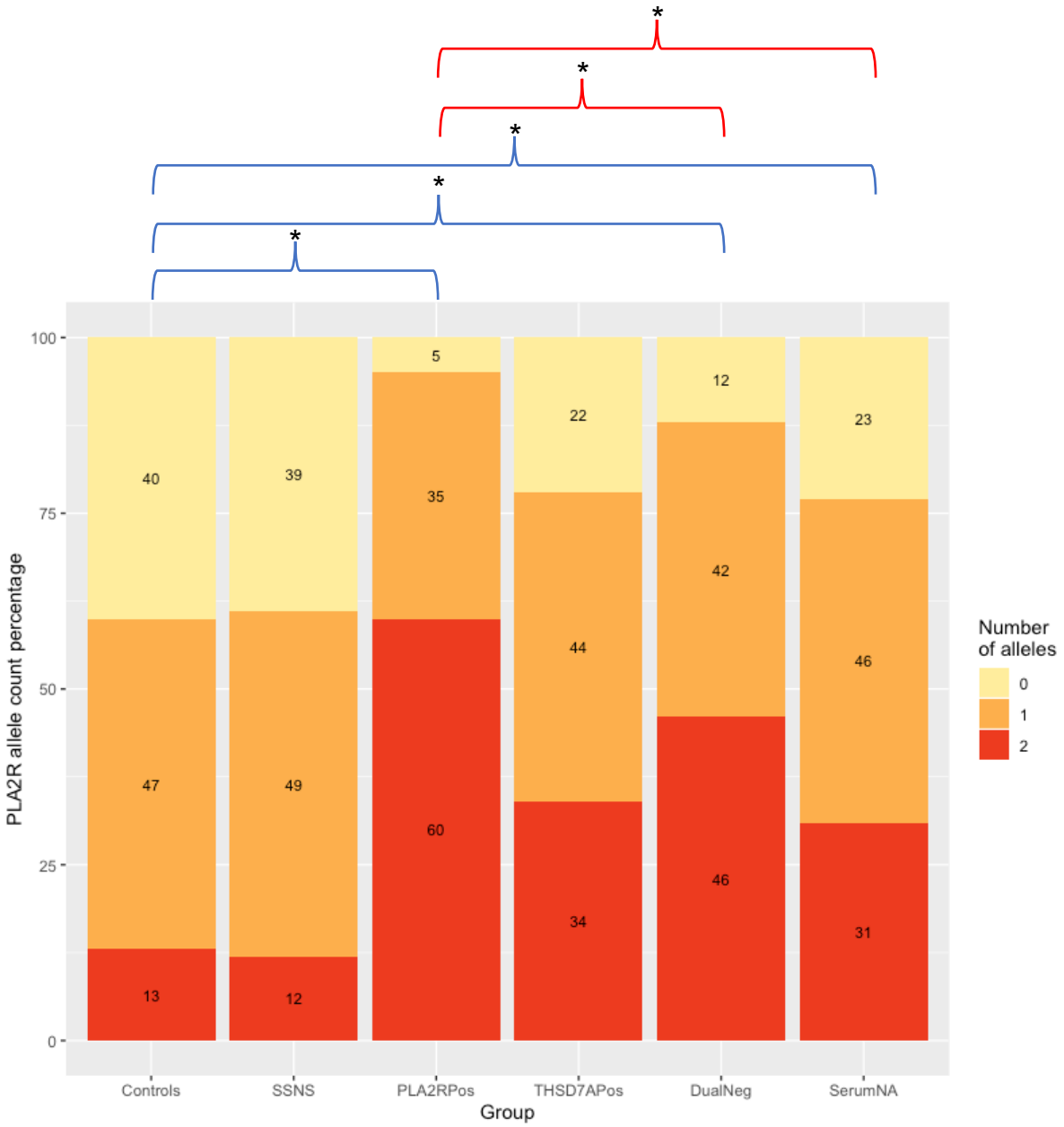


Figure 7.46: Stacked box plot of *PLA2R1* risk variant allele count proportions in different phenotype groups. The proportion of individuals within each group are demonstrated by the colour of the allele count stack. Statistically significant differences from Controls are highlighted by blue braces, (p <0.0001). Statistically significant differences from PLA2Rpos are highlighted by red braces, (p =0.0007 versus DualNeg and p =0.0003 versus SerumNA). For simplicity, statistical difference with SSNS is not shown but is the same as Controls, p <0.0001. SSNS, steroid sensitive nephrotic syndrome; PLA2RPos, anti-PLA2R1 antibody positive AMN; THSD7APos, anti-THSD7A antibody positive AMN; DualPos, PLA2R1 and THSD7A antibody positive AMN; DualNeg, PLA2R1 and THSD7A antibody negative AMN; SerumNA, serum not available for testing.

7.3.3.2. *HLA-DQA1* risk allele

The *HLA-DQA1* locus risk allele number demonstrated greater variability, Figure 7.47. Healthy subjects were statistically different to: the aPLA2Rab cases ($p < 0.0001$); dual antibody negative cases ($p < 0.0001$); and the serum unavailable group ($p < 0.0001$). The SSNS group was similar to healthy controls and similarly differed from the same three MN groups ($p < 0.0001$ in all groups). For this locus the anti-THSD7A antibody cases were statistically different to the aPLA2Rab cases ($p < 0.0001$), the dual antibody negative cases ($p = 0.0001$) and the serum unavailable group ($p < 0.0001$) but not controls.

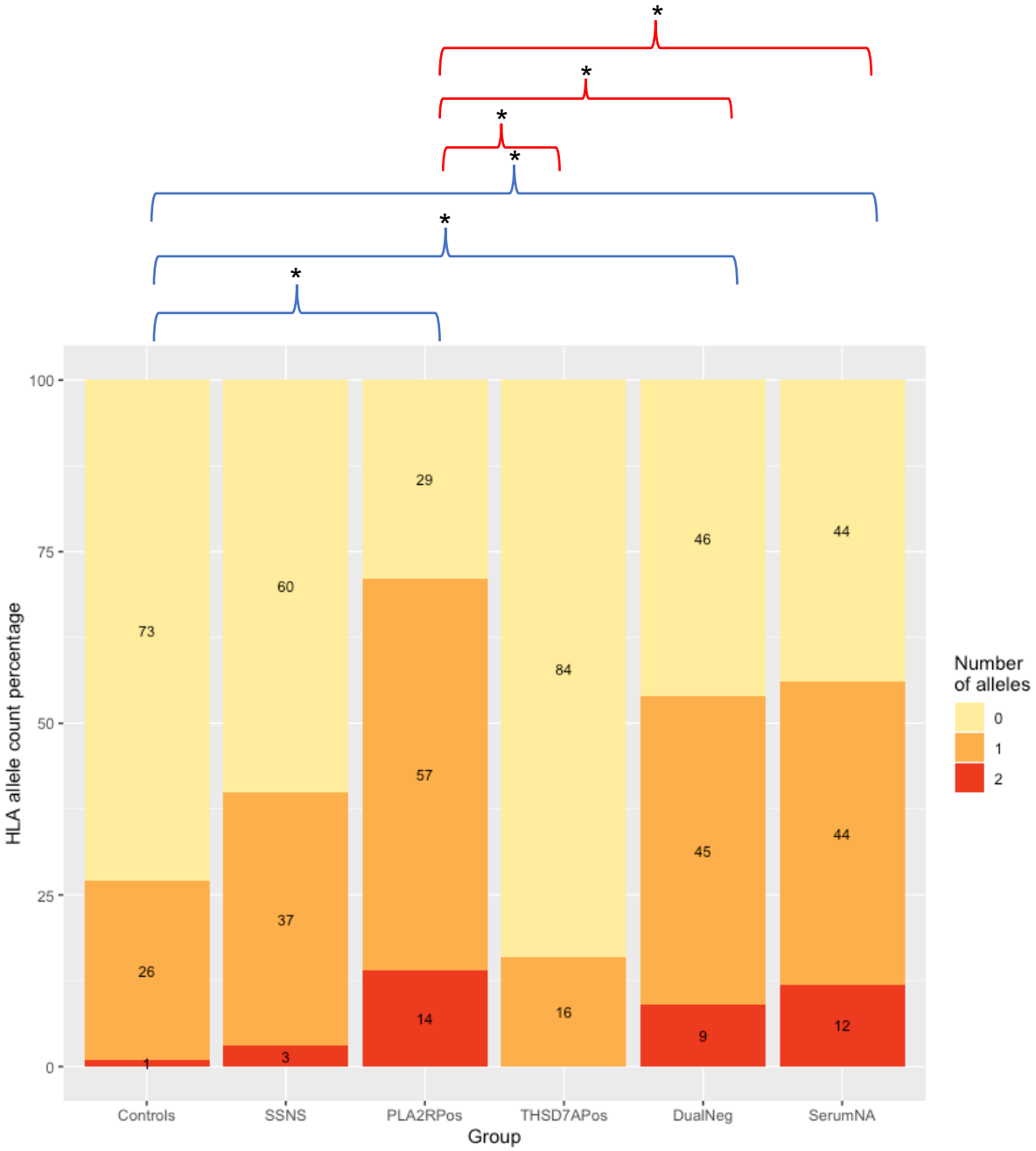


Figure 7.47: Stacked box plot of *HLA-DQA1* risk variant allele count proportions in different phenotype groups. The proportion of individuals within each group are demonstrated by the colour of the allele count stack. Statistically significant differences from Controls are highlighted by blue braces, (p <0.0001). Statistically significant differences from THSD7APos are highlighted by red braces, (p <0.0001 versus PLA2Rpos and SerumNA, p =0.0001 versus DualNeg). For simplicity, statistical difference with SSNS is not shown but is the same as Controls, p <0.0001.
SSNS, steroid sensitive nephrotic syndrome; PLA2RPos, anti-PLA2R1 antibody positive MN; THSD7APos, anti-THSD7A antibody positive MN; DualPos, PLA2R1 and THSD7A antibody positive MN; DualNeg, PLA2R1 and THSD7A antibody negative MN; SerumNA, serum not available for testing.

7.3.4. Age and genetic risk

Analysis of two subsets of the aPLA2Rab positive cases with demographic data was performed. GRS and age of onset of disease are inversely associated among both the 243 German and 117 British aPLA2Rab positive cases (p-value =0.005 and 0.02 respectively, Figure 7.48). Because the findings were replicated in the separate analyses the German and British datasets were combined to assess the effect of risk loci.

In the combined cohort of 342 European aPLA2Rab AMN linear regression with the *PLA2R1* risk allele showed no association, Figure 7.48c. However, the *HLA-DQA1* risk allele was associated with age (p = 0.001, Figure 7.48d). The higher the *HLA-DQA1* allele count the younger the age of onset.

7.3.5. Other clinical parameters

Among the 225 German subjects for whom uniformly collected phenotype data were available, regression analyses of GRS with uPCR, glomerular filtration rate decline per year (minimum five-year follow-up data) with anti-PLA2R1 antibody titre, immunosuppression, gender and age as co-variates revealed no statistically significant associations.

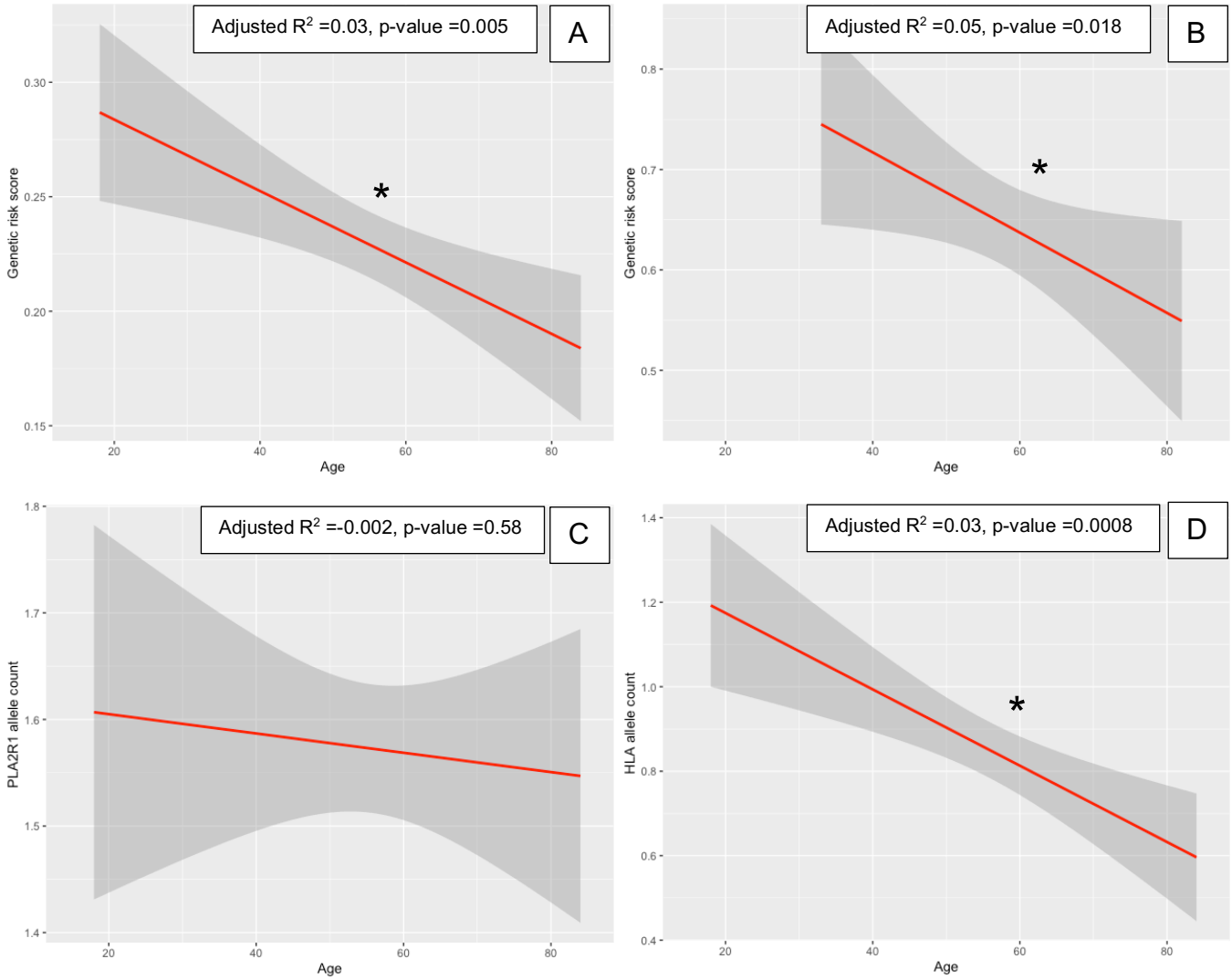


Figure 7.48: Linear regression of age of onset in anti-PLA2R1 antibody AMN, statistically significant graphs are marked by an asterisk with p-values shown in the box.
a. Linear regression of combined genetic risk score and the age of onset of anti-PLA2R1 antibody AMN in a German European cohort, n =225.
b. Linear regression of combined genetic risk score and the age of onset of anti-PLA2R1 antibody AMN in a British European cohort, n =117.
c. Linear regression of allele count in the *PLA2R1* locus and the age of onset of anti-PLA2R1 antibody AMN in the combined European cohort, n =342.
d. Linear regression of allele count in the *HLA-DQA1* locus and the age of onset of anti-PLA2R1 antibody AMN in the combined European cohort, n =342.

7.3.6. Paediatric onset AMN

A cohort of 15 individuals with paediatric onset of aPLA2Rab negative AMN was investigated. Due to the historical nature of this cohort Semaphorin 3B had not been discovered at the time of DNA and serum collection so this has not been measured, see 5.1.7.5.

Post genotyping there were 713,599 SNV markers for analysis. QC per individual removed 0 individuals. Per SNV filtering excluded 58,576 SNVs for call rate <95% and 107,747 SNVs for MAF <5%, leaving 547,276 SNV markers. From the filtered genotyping data, the two lead SNVs for *PLA2R1* and *HLA-DQA1* were extracted.

First the paediatric AMN group were compared to the aPLA2Rab AMN cases. A Wilcoxon rank sum test demonstrated that there were statistical differences in all three components between these two groups; the GRS ($p = 0.002$), the *PLA2R1* allele count ($p = 0.0007$) and *HLA-DQA1* allele count ($p = 0.03$), Figure 7.49.

I then decided to compare the paediatric AMN cases to healthy controls and SSNS in addition to aPLA2Rab positive AMN cases. A Dunn's test for multiple groups testing demonstrated a statistically significant difference ($p < 0.02$) between the paediatric AMN group and the aPLA2Rab positive group in all three domains; GRS ($p = 0.09$), *PLA2R1* allele count ($p = 0.004$) and *HLA-DQA1* allele count ($p = 0.02$), Figure 7.50 and Figure 7.51 respectively. There were no further statistically significant associations between the paediatric group and either controls or the SSNS individuals.

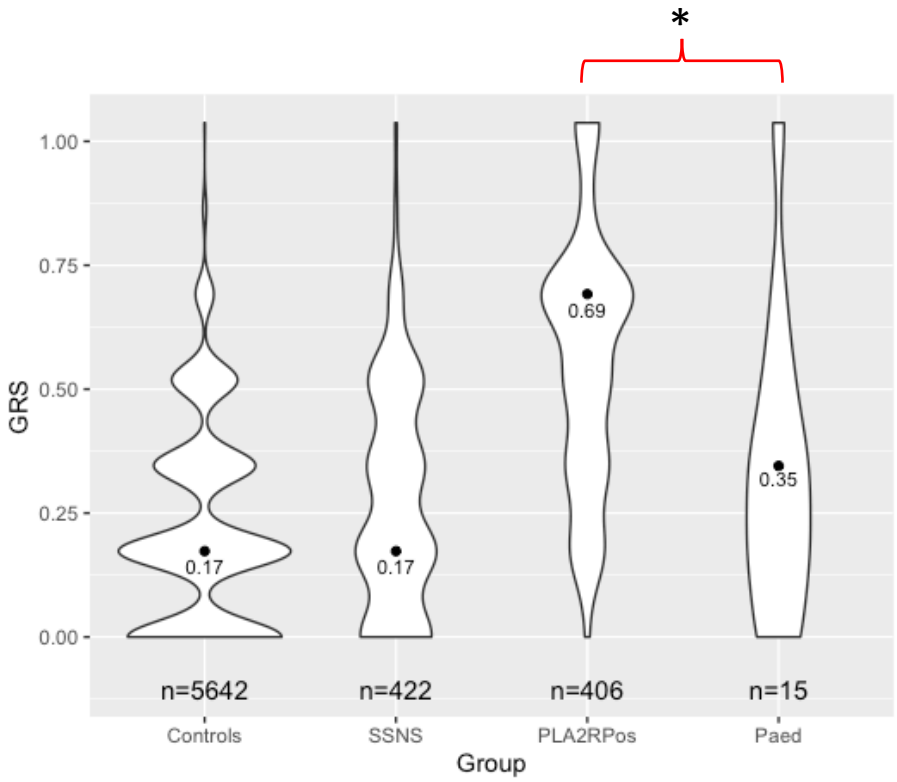


Figure 7.49: Violin plot of genetic risk score in different phenotype groups. Number of individuals in each group is shown. The black dot represents the median. SSNS, steroid sensitive nephrotic syndrome; PLA2RPos, anti-PLA2R1 antibody positive AMN; Paed, Paediatric AMN.

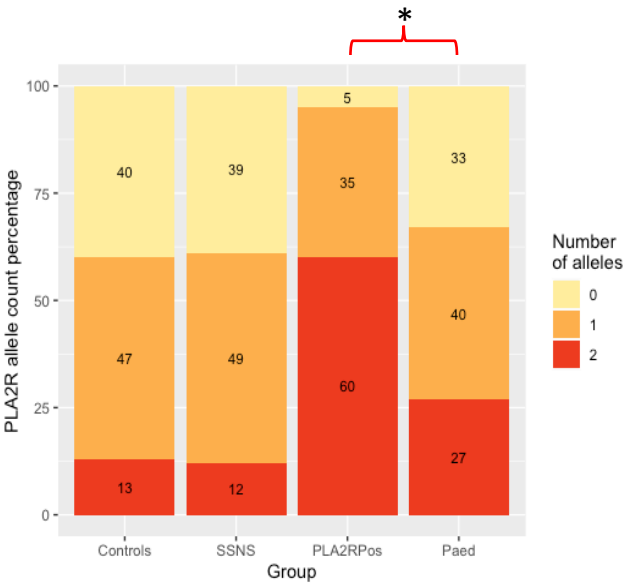


Figure 7.50: Stacked box plot of *PLA2R1* risk variant allele count proportions in different phenotype groups. The proportion of individuals within each group are demonstrated by the colour of the allele count stack. Statistically significant differences are highlighted by red braces, (p <0.05). SSNS, steroid sensitive nephrotic syndrome; PLA2RPos, anti-PLA2R1 antibody positive AMN; Paed, paediatric AMN.

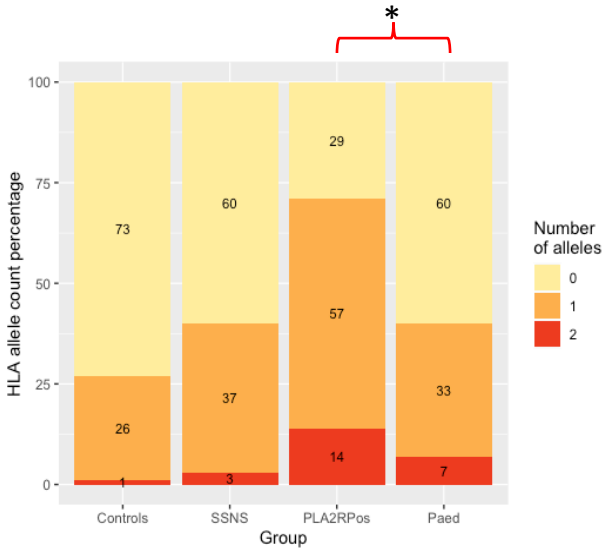


Figure 7.51: Stacked box plot of *HLA-DQA1* risk variant allele count proportions in different phenotype groups. The proportion of individuals within each group are demonstrated by the colour of the allele count stack. SSNS, steroid sensitive nephrotic syndrome; PLA2RPos, anti-PLA2R1 antibody positive AMN; Paed, paediatric AMN.

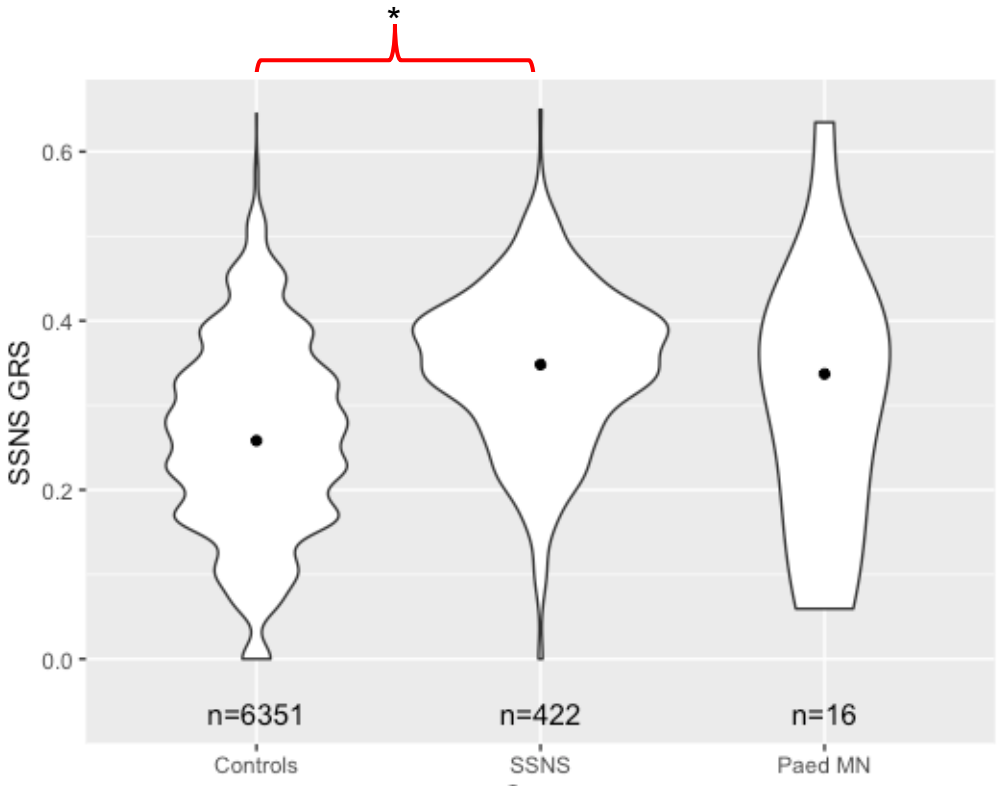


Figure 7.52: Violin plot of the SSNS genetic risk score in different phenotype groups. Number of individuals in each group is shown. The black dot represents the median. Statistically significant differences are highlighted (p <0.05). SSNS, steroid sensitive nephrotic syndrome; Paed MN, paediatric AMN.

Even though the paediatric AMN cases were biopsy-proven cases as an additional check because of the overlap with SSNS in paediatric nephrotic syndrome I calculated the SSNS genetic risk score in controls, individuals with SSNS and the paediatric AMN cases, Figure 7.52. This demonstrated that there was a clear difference between SSNS and healthy controls (p =0), no difference between the paediatric AMN cases and controls (p =0.38) and a trend to difference between SSNS and paediatric AMN (p =0.09).

7.3.7. Anti-contactin-1 antibody associated AMN

A cohort of 7 individuals with anti-contactin-1 antibody associated AMN with CIDP was investigated. No individuals were removed with per individual QC and per SNV QC matches that as summarised in Figure 7.22. Ancestry was examined through a principal components analysis, see Figure 7.53. This identified 1 individual did not cluster around the European reference controls and were more likely of African ancestry, so they were excluded, highlighted with the red circle in Figure 7.53, resulting in 6 individuals for analysis. From the filtered genotyping data, the two lead SNVs for *PLA2R1* and *HLA-DQA1* were extracted.

The GRS in the anti-contactin antibody AMN cases is statistically different to the aPLA2Rab AMN cases ($p = 0.01$) but not the controls ($p = 0.93$), Figure 7.55. The *PLA2R1* allele count was statistically different between the anti-contactin antibody and aPLA2Rab positive cases ($p = 0.0004$) but not with controls ($p = 0.61$), Figure 7.54. Interestingly, there was no statistically significant difference with the *HLA-DQA1* allele count when compared to either the aPLA2Rab group ($p = 0.22$) or the controls ($p = 0.22$), Figure 7.56.

As an additional stringent statistical test, a Dunn's test for multiple testing was done and this confirmed the same findings; the anti-contactin antibody group are statistically different to the aPLA2Rab positive cases in the GRS and the *PLA2R1* risk variant allele count but not the *HLA-DQA1* allele count.

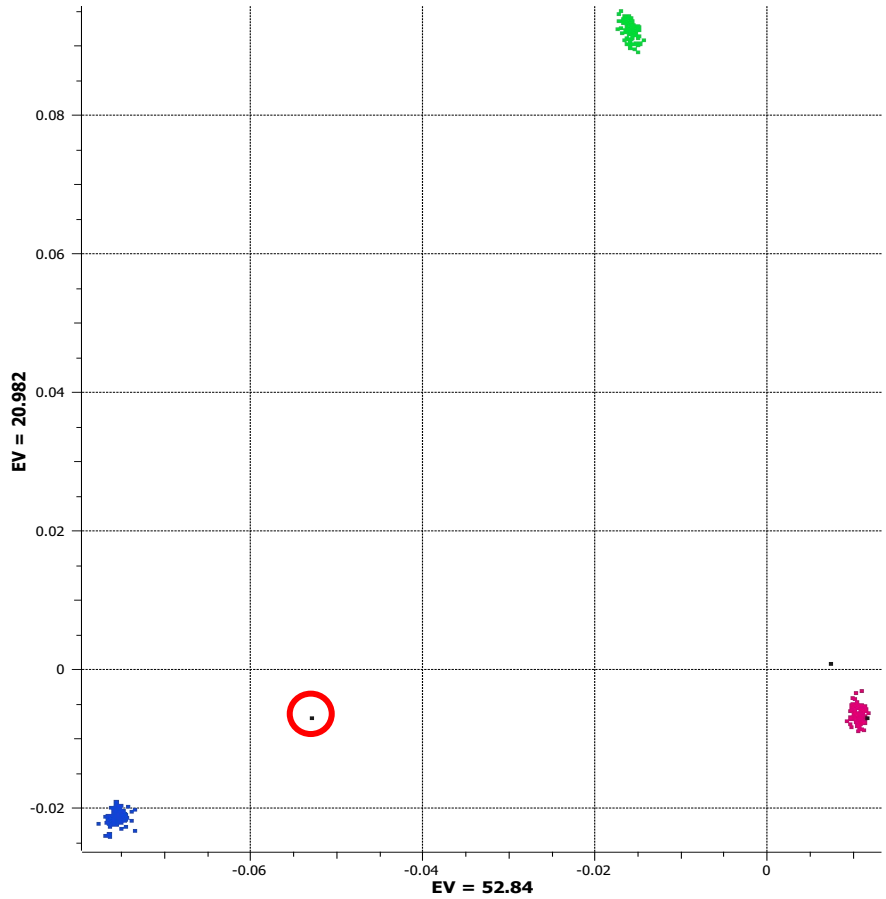


Figure 7.53: Principal components analysis to demonstrate divergent ancestries of the anti-contactin-1 antibody MN cases with the 1000 Genomes Project reference controls. Cases are in black, European ancestry controls in pink, African ancestry controls in blue and East Asian ancestry controls in green. The 1 outlier is highlighted with the red circle.

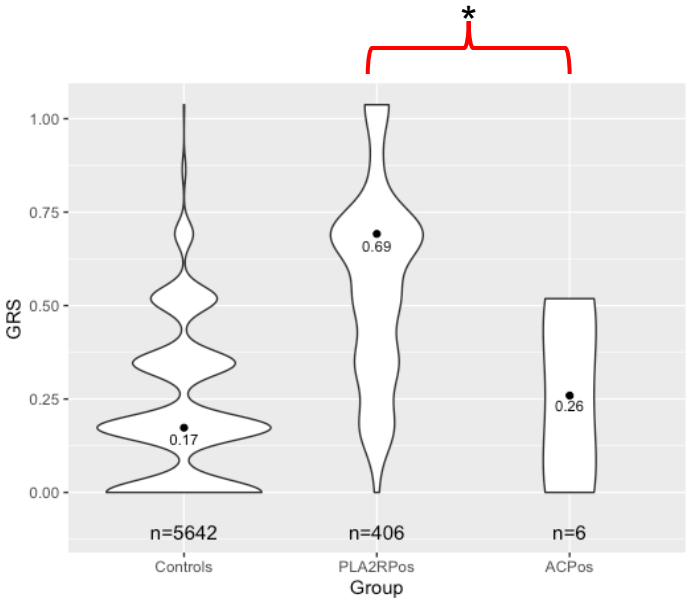


Figure 7.55: Violin plot of genetic risk score in different phenotype groups. Number of individuals in each group is shown. Statistically significant differences are highlighted by red braces, (p <0.05) but are not shown for controls versus aPLA2Rab differences. The black dot represents the median. PLA2RPos, anti-PLA2R1 antibody positive AMN; ACPos, anti-contactin antibody positive AMN.

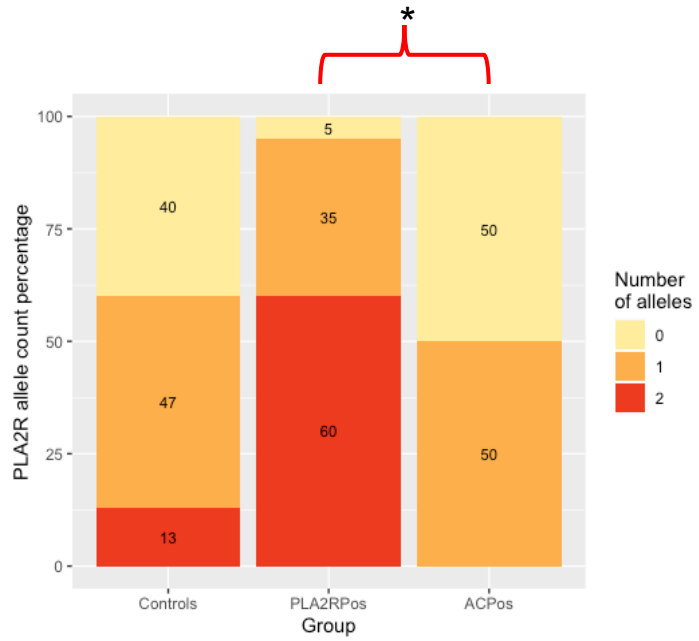


Figure 7.54: Stacked box plot of *PLA2R1* risk variant allele count proportions in different phenotype groups. The proportion of individuals within each group are demonstrated by the colour of the allele count stack. Statistically significant differences are highlighted by red braces, (p <0.05) but are not shown for controls versus aPLA2Rab differences. PLA2RPos, anti-PLA2R1 antibody positive AMN; ACPos, anti-contactin antibody positive AMN.

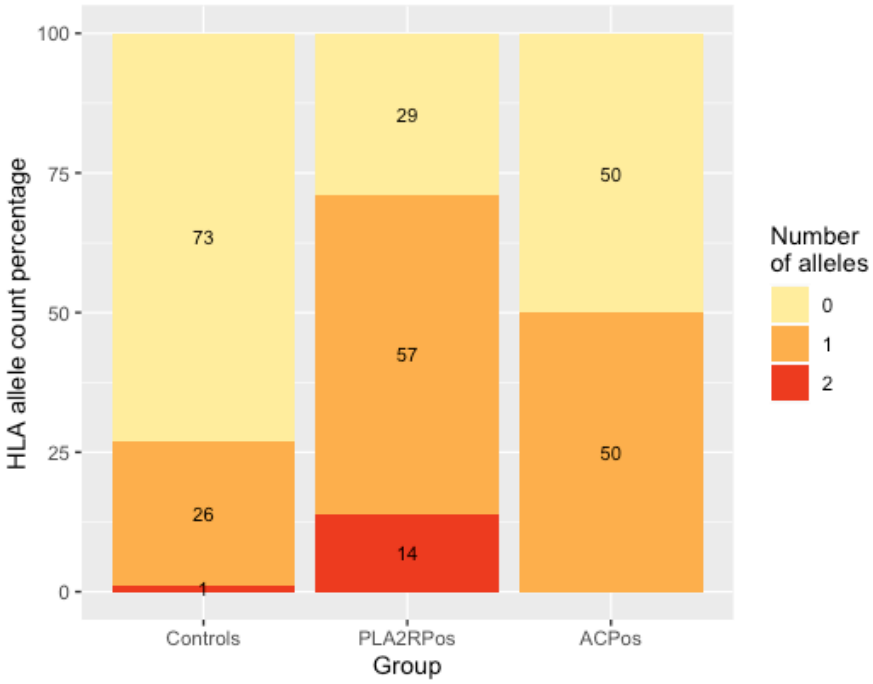


Figure 7.56: Stacked box plot of *HLA-DQA1* risk variant allele count proportions in different phenotype groups. The proportion of individuals within each group are demonstrated by the colour of the allele count stack. PLA2RPos, anti-PLA2R1 antibody positive AMN; ACPos, anti-contactin antibody positive AMN.

7.4. UK Biobank

7.4.1. Quality control

There were 488,377 individuals with 784,256 SNVs available for analysis. Standard QC protocols as per 6.2.5 in PLINK were done.

7.4.1.1. Per individual

The per individual QC excluded 11,065 individuals for heterozygosity rate >3 SD. The genotyping call rate per individual was low and using a stringent criteria of call rate >98% excluded all individuals, Figure 7.57. For this reason, I decided to use less stringent criteria and instead used a call rate of >90% which did not exclude any individuals.

The IBD calculations in UKBB were tried with two different parameters. UKBB state that there are 107,162 related pairs of individuals within the dataset using a more stringent IBD co-efficient score of <0.1768 to exclude greater than third degree individuals [352]. I calculated the IBD with the KING co-efficient tool (same tool as UKBB) and for an IBD score of <0.1768 only excluded 34,691 individuals, which is a difference of 18,890 individuals. With a less stringent IBD score <0.1875 this excluded 34,255 individuals. Because I wanted to include as many individuals as possible and there was already a discrepancy, I decided to proceed with the less stringent criteria for IBD (<0.1875); this left 443,063 individuals for analysis.

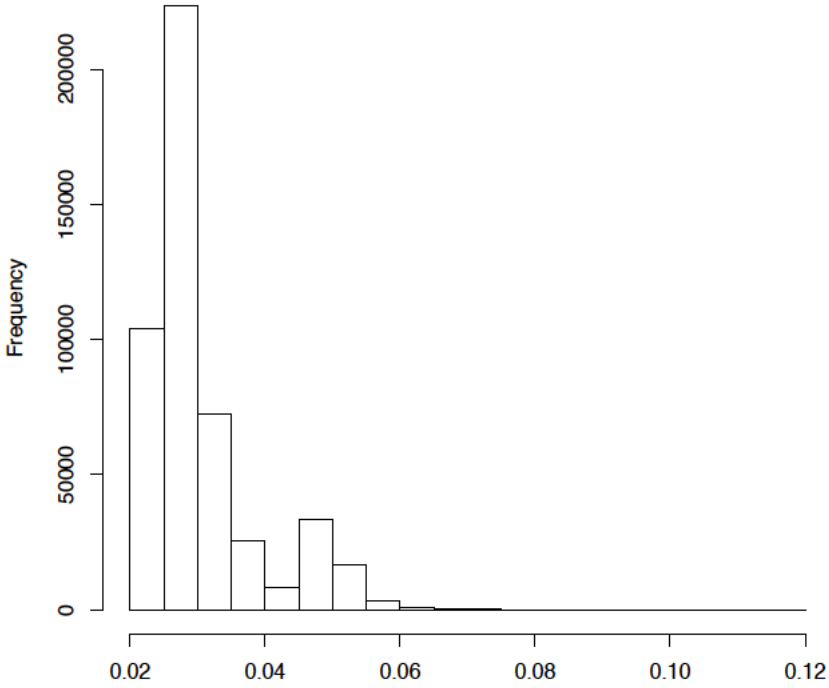


Figure 7.57: Histogram of individual genotyping missingness rate. The x-axis is the proportion of missing genotypes per individual.

7.4.1.2. Per SNV

The per SNV QC excluded; call rate <98% 102,326, Figure 7.58; minor allele frequency <1% 98,020, Figure 7.59; HWE <0.001 212,341. No information on the X-chromosome was available and non-biallelic SNVs had already been filtered out. This left 371,569 SNVs for further analysis.

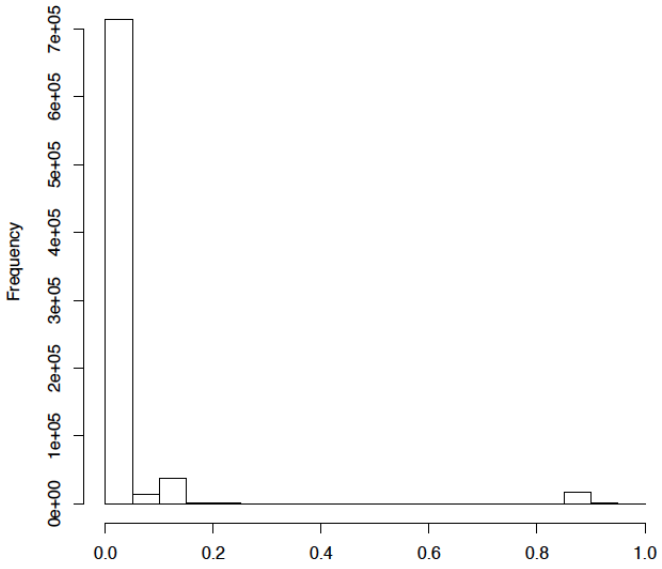


Figure 7.59: Histogram of SNV genotyping call rate. The x-axis is the proportion of missing genotypes per SNV.

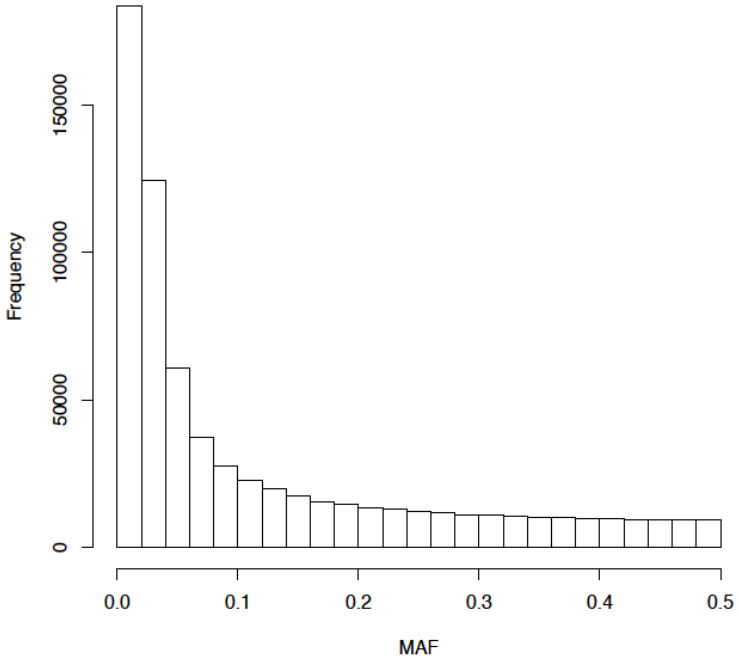


Figure 7.58: Histogram of the distribution of the minor allele frequency (MAF) of all SNVs.

7.4.1.3. Population stratification

The GRS score is only valid in those of European ancestry and so these individuals needed to be extracted from the UKBB. The method of principal component analysis was attempted to undertake this.

The post QC UKBB data was merged with the 1000 Genomes project dataset. There were 227,967 intersecting SNVs and these were pruned first to 39,815 independent SNVs. To facilitate computational processing time and analysis I decided to prune the data further to 10,000 SNVs. There were 443,063 UKBB individuals and 629 1000 Genomes Project individuals. The principal components are shown in Figure 7.60.

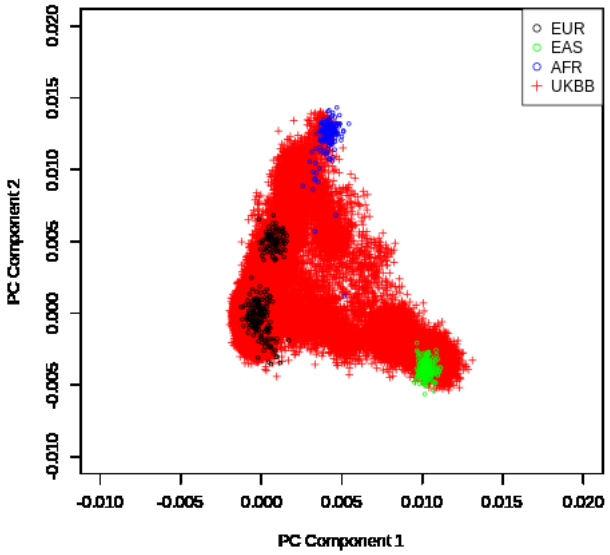


Figure 7.60: Principal component analysis of divergent ancestries in UKBB and reference controls. UKBB controls are highlighted in red, European ancestry controls in black, African ancestry controls in blue and East Asian ancestry controls in green.

Due to the size of the dataset the usual tools used to extract only the European individuals were not able to process a dataset of this size. I attempted to use SVS which did not have the appropriate computational power, SmartPCA estimated that the analysis would take 51 years and the other program I tried was called bigsnpr which did not have all the code readily available [334, 353].

Other published studies utilising the UKBB dataset use self-reported ancestry information. Without any other suitable options, I too, decided to extract the UKBB self-reported ancestry individuals; 472,725 are of European ancestry. To visualise the diversity within this group I undertook a PCA, Figure 7.61.

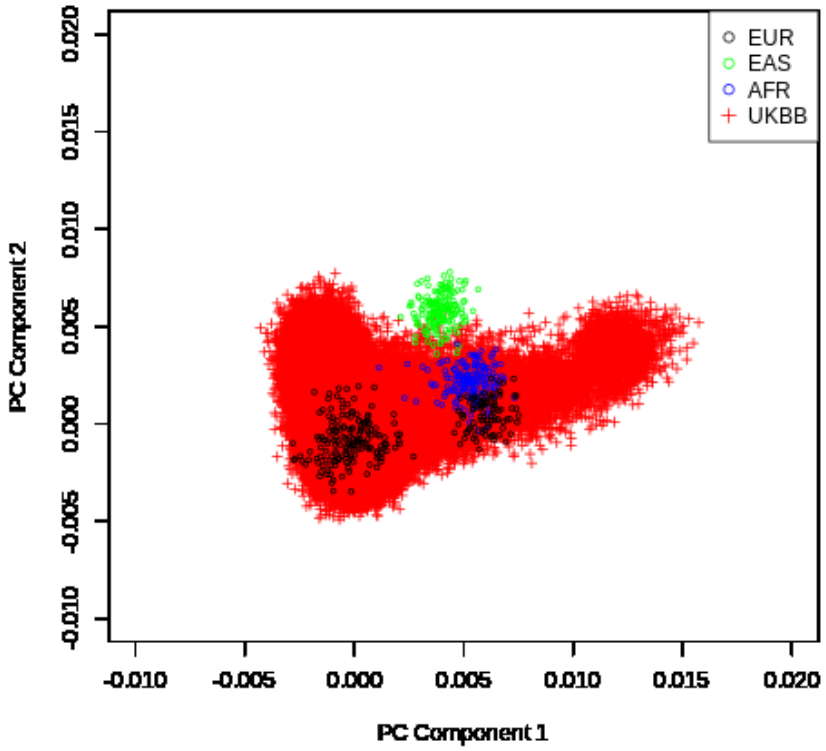


Figure 7.61: Principal component analysis of self-reported European individuals from the UKBB and ancestry reference controls. UKBB European individuals are highlighted in red, European ancestry controls in black, African ancestry controls in blue and East Asian ancestry controls in green.

7.4.2. GRS in UK population

The GRS was calculated in the self-reported white European ancestry individuals. To facilitate a uniform dataset only individuals that had both risk SNVs genotyped were examined and the others were excluded from analysis. For the analysis there were 419,802 European individuals from the UKBB with data in both risk SNVs for AMN.

The number and proportion of individuals with the different numbers of *PLA2R1* risk alleles and *HLA-DQA1* risk alleles is shown in Table 7.18. The GRS was calculated and the proportion of individuals from the population at risk of AMN having all 4 risk alleles is 0.8%, Table 7.19 and Figure 7.62. A comparison of the distributions between the GRS from healthy controls and confirmed aPLA2Rab positive cases was done with the UKBB dataset, Figure 7.62.

No of *PLA2R1* risk allele	**No of individuals**	**Percentage of individuals**
0	73,384	17.48%
1	204,003	48.6%
2	142,415	33.92%
No of HLA-DQA1 risk allele	**No of individuals**	**Percentage of individuals**
0	307,868	73%
1	101,918	24.28%
2	10,016	2.39%

Table 7.18: Table showing the percentage of European individuals from UKBB with the number of risk alleles for AMN

Genetic risk score	**No of individuals**	**Percentage of individuals**
0	53,780	12.81%
0.173	149,430	35.6%
0.345	17,841	4.24%
0.347	104,658	24.93%
0.519	49,692	11.84%
0.691	1,763	0.42%
0.692	34,385	8.19%
0.864	4,881	1.16%
1.04	3,372	0.8%

Table 7.19: Table showing the percentage of European individuals from UKBB with the number of risk alleles for AMN

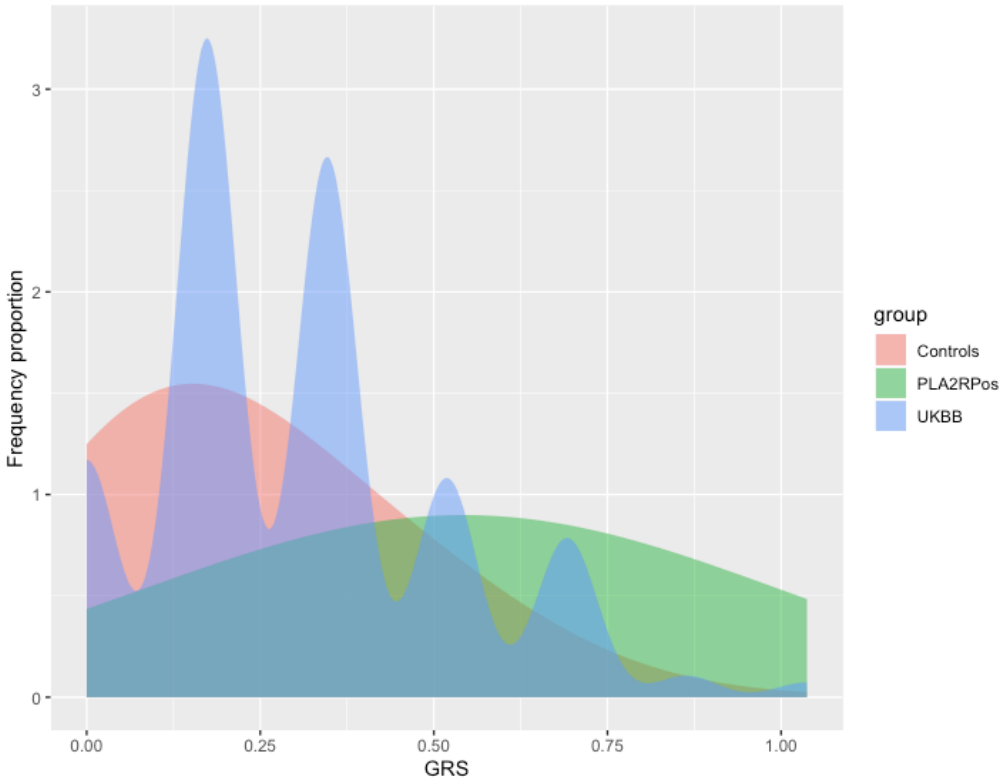


Figure 7.62: Histogram showing the genetic risk score distribution across the three different cohorts. Controls = healthy European controls, PLA2RPos = aPLA2Rab positive European AMN cases, UKBB = self-reported European individuals from UK Biobank.

While the proportion of individuals having the risk allele in *PLA2R1* is high when this is combined with the *HLA-DQA1* risk allele it becomes rare and less frequent.

Due to the apparent skewing of data for the *HLA-DQA1* risk allele count towards a lower allele count I did a manual check to make sure that this SNV was not out of Hardy-Weinberg equilibrium:

HLA-DQA1 SNV
p^2 frequency = 0.73
2pq frequency = 0.24
q^2 frequency = 0.024
1 -0.73 -0.24 -0.024 = 0.03

This demonstrates that the HLA-DQA1 SNV does indeed follow the HWE criteria. Which means that the data analysis is reliable.

7.4.3. Further work

After the thesis submission I undertook some further work on the UKBB dataset. Based on hospital episode statistic data individuals with a diagnosis of membranous nephropathy were extracted. I calculated the allele counts for the first reported GWAS [73] and for the most recent GWAS [171]. From this data I calculated the relative risk for each genotype for having AMN. Colleagues did the same using the SNVs from a third GWAS [147]. This data is summarised Table 7.20. Colleagues further calculated the relative risk for AMN is 23.44 in the highest risk genotype, Table 7.21.

	SNV	Allele	AMN	RR (95% CI)
1	*HLADQA1* rs2187668	CC	44/357516 (0.01%)	1
		CT	39/118070 (0.03%)	2.68 (1.74-4.13)
		TT	14/11622 (0.12%)	9.79 (5.36-17.85)
	PLA2R1 rs4664308	GG	11/81861 (0.01%)	1
		AG	34/230759 (0.02%)	1.10 (0.56-2.16)
		AA	52/174588 (0.03%)	2.22 (1.16-4.25)
2	*HLADRB1/ DQA1* rs9271573	CC	24/170192 (0.01%)	1
		AC	41/234411 (0.02%)	1.24 (0.75-2.05)
		AA	32/82375 (0.04%)	2.75 (1.62-4.68)
	PLA2R1 rs17831251	TT	11/80771 (0.014%)	1
		CT	33/227843 (0.01%)	1.06 (0.54-2.10)
		CC	51/172495 (0.03%)	2.17 (1.13-4.16)
	NFKB1 rs230540	CC	11/59492 (0.02%)	1
		CT	43/216576 (0.02%)	1.07 (0.55-2.08)
		TT	41/205265 (0.02%)	1.08 (0.56-2.10)
	IRF4 rs9405192	AA	4/33043 (0.01%)	1
		GA	39/182247 (0.02%)	1.77 (0.63-4.95)
		GG	54/259096 (0.02%)	1.72 (0.62-4.75)

Table 7.20: Genetic risk analysis using the lead SNVs from GWAS 1 [73] and GWAS 2 [171] in the UK Biobank. SNV = single nucleotide polymorphism, AMN = autoimmune membranous nephropathy, RR = relative risk, CI = confidence interval.

Risk groups	Genotype	AMN	RR (95% CI)
Low Risk	CCGG	5/59756 (0.01%)	1
Medium Risk	All else	84/423373 (0.02%)	2.37 (0.96-5.85)
High Risk	TTAA	8/4079 (0.20%)	23.44 (7.67-71.62)

Table 7.21: Genetic risk and calculated relative risk of AMN (autoimmune membranous nephropathy).

8. Discussion

8.1. PLA2R1 intronic variant analysis

This work has identified intronic and exonic variants that are strongly associated with AMN in European patients. After quality control filtering of the initial 2203 variants and Chi-squared analysis a statistically significant association with 109 variants was found. Of the variants that had a statistically significant association only 2 were coding variants, these were the two previously described common exonic variants [72]. 7 variants were annotated to be in putative TFBS from the UCSC annotations although, further assessment revealed that none of these were within the specific transcription factor DNA binding domain. Association does not equal causation and unidentified confounders could be present that could affect both AMN and the identified variants [354]. The association itself does not mean direct causality, however in the absence of any other information it suggests a hypothesis [354]. The stronger the association the stronger the potential hypothesis, however the more variables that are examined the more likely spurious associations will be made [354]. It was for this reason that I decided to focus the initial examination of the potential association with AMN on the lead variant with the caveat in mind that correlation does not amount to causation.

8.1.1. Lead variant

The power of association of the most statistically significant associated variant, rs528521365, is exceptionally large with a p-value of 7.96×10^{-227}. This variant is present in 98% of alleles in 334 AMN patients compared to only 17% of controls. It

was because of the strength of this association that I sought to investigate if there was a potential causative mechanism. I attempted to ascertain the potential impact of this variant and potential mechanism of inducing disease. The initial UCSC annotation suggested that this variant was in a TFBS which is why it was flagged in the analysis, however, it is 66 bp downstream of a TFBS for CEBPB. The c-Fos, c-Jun and CEBPB complex starts at this position so this may be implicated.

8.1.1.1. CCAAT/enhancer binding protein beta

CEBPB regulates expression of genes involved in immune and inflammatory responses [59, 355-357] which provided further support to its potential causative role in AMN; a disease driven by an abnormal immune response. Further both CEBPB and PLA2R1 have been identified and localised to alveolar epithelial cells; in response to the respiratory syncytial virus, CEBPB is synthesised rapidly in pulmonary alveolar epithelial cells within 3 hours of infection and continues to accumulate until 48 hours [358]. Similarly, PLA2R1 has been found to be overexpressed in alveolar epithelial brushings of asthmatic children and *PLA2R1* knockout mice have more dendritic cells, increased levels of IgG and increased production of cytokines by lung leucocytes [359, 360]. The presence of both proteins co-localising to the same tissue cells [358] and involvement of the immune system with CEBPB in respiratory diseases [358] provided further support to the potential role of CEBPB in AMN. CEBPB has been identified as an important transcription factor regulating expression of genes involved in both immune and inflammatory responses [361].

The contrasting evidence to this is that to date there is no published evidence of a renal phenotype in *CEBPB* knockout mice. *CEBPB* knockout mice are small, have reduced bone mass and are protected from obesity [362-364]. However, CEBPB expression was increased in animal models during worsening kidney injury and could be a mechanism of kidney damage [365]. This provided further implications of the role of CEBPB expression in a diseased kidney to support my follow up experiments.

A specific and potential role in AMN can be proposed as CEBPB has an anti-proliferative effect via Myc on T-cells and facilitates T-helper cell 2 (Th2) differentiation [59, 355-357]. The most active subclass in aPLA2Rab mediated AMN is IgG4 [366]. It is this very IgG subclass (IgG4) that is upregulated during Th2 responses and has been found to be specific to AMN compared to other glomerulonephritic diseases [367]. There is no evidence to date to suggest an interaction between CEBPB and PLA2R1, however, it is possible CEBPB could facilitate Th2 differentiation in AMN. A variant in the CEBPB TFBS in *PLA2R1* may disrupt this binding; either by strengthening the binding to create a prolonged activation of the Th2 response by CEBPB or by weakening the binding - this is speculative.

8.1.1.2. Functionality of proposed variant

To investigate the possible functionality of the lead variant in *CEBPB* I considered bioinformatic assessment for *in silico* predictions and *in vitro* experiments. *In vivo* experiments are difficult as the variant and the nearby genomic region is not conserved across species and in particular is not present in commonly used animal

models such as mice, rats, rabbits or pigs [70]. Rat models have been established with Heymann's nephritis to model membranous nephropathy but these are not helpful [368].

8.1.1.2.1. Bioinformatic assessment of function

8.1.1.2.1.1. CEBPB motif

The CEBPB DNA binding motif was confirmed with MAST because the ENCODE consortia ChIP-Seq experiments about CEBPB is derived from a single submitted experiment so is not entirely reliable [369, 370]. ENCODE have performed ChIP-seq experiments on only 140 different transcription factors and histone modifications [339] and is not a fully comprehensive list of all identified human transcription factors. Depending on the database used, the estimates for the total number of transcription factors ranges from approximately 2000 to 3000. UniProt estimates 1317, Gene Ontology 2718 and the DNA Binding Domain transcription factor database 2886 [371-374], which is a large disparity. It is difficult to assess TFBSs if there are discrepancies as to what transcription factors exist. There are limitations within the ENCODE data and these provide further explanation for the vast discrepancy in identified transcription factors and confirmed binding sites. One of the biggest limitations with ChIP-seq is the lack of available antibodies for the transcription factors with subsequent lack of enrichment in the affinity precipitation step [340]. The ENCODE consortia state that they have the lowest confidence about data on the function and binding of a single transcription factor supported by ChIP-seq [375]. Another factor may be that the data on transcription factor occupancy is dependent on p-values which are set conservatively and are dependent on statistical analyses [375]. These factors meant that the CEBPB binding motif results may not be as

reliable as first predicted and more importantly that other transcription factors may bind here.

8.1.1.2.1.2. Alternative transcription factor binding sites

There are multiple search tools for predicting TFBS and these yield different results. Utilising Patch 1.0, the control population with the allele A have a TFBS starting at the very position of the variant - 289 - with the sequence TGACGTAGT on the reverse strand for the the c-Fos, c-Jun and CEBPB complex. The AMN variant reduces the binding score and may potentially reduce the affinity of binding. c-Fos and c-Jun form a heterodimer to create the AP-1 transcription factor complex and subsequently is involved in TGF-beta mediated signalling [376]. This AP-1 transcription factor complex interacts with the immune system making it a potentially implicated transcription factor. The unusual aspect is that the proposed mechanism of the inability to bind in AMN precipitates disease whereas one would expect increased binding increases the Th2 response. This is potentially the most promising and interesting prediction to suggest an alternative TFBS within the region of the variant which has the potential to be pathogenic.

As there are up to 2746 transcription factors that have not been examined by ENCODE it is possible that one of these transcription factors may bind to the region near the lead variant. This was investigated with the TRANSFAC database and this identified another potential binding sites for Pax-6 in humans. Pax-6 is not expressed in the kidney and is associated with neural development, in particular oculogenesis [377, 378], so did not seem relevant. Alibaba implicates Sp1 which may be implicated as it promotes activity of podocalyxin in rat podocytes [379]. Sp1 supports

transcriptional regulation of podocalyxin which is an integral membrane protein in the foot processes of the rat podocytes that helps maintain the interdigitations of the podocyte foot for urinary filtration [379]. It is possible that Sp1 binding affects the podocalyxin protein function, causing podocyte disease or AMN.

The different results and implicated transcription factors based on these *in silico* analyses demonstrate the difficulty with these techniques. They are still viewed as hypothetical based on algorithms and calculations; so functional *in vivo* experiments were necessary.

8.1.1.2.2. Experimental assessment of variant function

Laboratory based techniques such as *in vitro*, *ex vivo* or *in vivo* are considered to be reliable methods of validating and proving the *in silico* analyses due to their historical nature, accepted guidelines and techniques. The lead variant is intronic so the direct effect on protein expression is not immediately demonstrable. Unfortunately, the variant is not in a conserved region in animals and so animal models are not amenable. It is commonly accepted that conserved regions are functionally important [380] so it was surprising that this region in *PLA2R1* was not. However, there is evidence that non-conserved regions can have a role in gene regulation and in enhancer regions [381, 382]. For this reason, I continued to investigate the lead variant. The two common mechanisms that intronic variants can have an affect are either with alternative splicing of the mRNA or affect protein binding for repressors or enhancers.

8.1.1.2.2.1. Transcription factor binding site

Although a few different transcription factors may be implicated with this variant I only investigated the role of CEBPB with EMSA. EMSA did not demonstrate any difference in binding of the transcription factor to DNA with or without the variant. While this was repeated three times, EMSAs do not often meet the required throughput for transcription factor and DNA binding site interactions [383]. One problem may be that the TNT mix system produces additional proteins (visible on the Western blot) that may interact and/or affect DNA binding. One way to overcome this may be to extract CEBPB from nuclear extracts.

Alternative methods are available and these could have been pursued. DNA immunoprecipitation (DNA-chip) utilises purified chromosomal DNA instead of synthesised DNA which is immunoprecipitated with protein-specific antibodies [383]. The formalin-fixed paraffin-embedded (FFPE) AMN renal biopsy tissue could be used for an *ex vivo* method such as ChIP-seq. This has a higher throughput however is dependent on protein abundance, cross-linking efficiency and antibody availability and specificity [383]. Proteomics of isolated chromatin segments (PICh) investigates associated proteins to a genomic locus using mass spectrometry and has the advantage of being indiscriminate about indirect or direct transcription factor binding to DNA [383].

8.1.1.2.2.2. Alternative mRNA splicing

The computational prediction tools did not predict an effect on alternative splicing between AMN and the controls however these are not completely reliable. It is not known if differences occur between PLA2R1 isoform expression in AMN but Western

blotting first identified PLA2R1 in glomeruli from AMN patients with controls missing the same protein band [57]. This suggests that protein expression is different in AMN glomeruli. *PLA2R1* messenger RNA is directly correlated with aPLA2Rab titres being negative in remission and higher in active disease [384]. Renal biopsies are a perfect source to obtain RNA as it is the tissue that is directly affected. However, renal biopsies are stored in FFPE which can make RNA extraction more difficult due to degradation [385]. RNA transcripts can be studied with reverse transcription PCR followed by quantitative PCR [369]. There are technologies that can improve RNA extraction from FFPE renal biopsies such as RNAse scope which utilises *in situ* hybridisation to RNA [386].

8.1.1.3. Reliability of lead variant

The biggest concern that I had throughout the analysis of the lead variant was the low functional impact score. deCODE provide a 'SNV score' for each variant in the output results. This is a log likelihood ratio test score, compared to a chi square table with one degree of freedom. deCODE state, "The higher the SNV score the more likely the SNV is real. We only report SNVs with a score of 10 or higher. Real SNVs often get a score of several hundred, even several thousand". My concern was that the lead variant's score was only 13, and therefore I did not think it was truly reliable. However, this was a deCODE quality control measure cut-off which had to be trusted to avoid a biased selection of results and so I pursued the investigation with the lead variant.

8.1.2. Replication

Replication was not successful despite pursuing and investigating multiple different methods. I have access to DNA from 1409 European AMN patients for a replication study. KASP genotyping failed due to the high GC content and polymorphic nature of the lead variant. LGC Genomics found that the C allele was amplifying on their control samples even though this is present in only 17% of the controls. This is the same allele that was amplified and sequenced in our original sequencing and so suggests there is an amplification bias.

The PCR products are of the correct size suggesting that the DNA amplification worked. There can be different reasons for the failure of Sanger sequencing. This could be due to insufficient DNA; I sent 10ng of DNA as the overall concentration from the purified PCR product was low. The purified PCR product DNA may be a poor product due to the purification process which can contain salts and buffers that can interfere or destroy the DNA [387]. To overcome this as a potential problem ExoSAP-IT is an alternative PCR product purifier that could be tried [388]. There are other suggested problems which have been excluded: loss of sequencing products during the clean-up, wrong primer use, impure water, primer degradation, degraded primers [387]. Problems beyond my control due to outsourcing to GATC are loss of sequencing products during the clean-up, degradation of *Taq* DNA polymerase and ddNTPs and a blocked capillary [387].

A major issue is the high numbers of repeats and polymorphisms surrounding this variant. As a result, sequencing by standard techniques may not be possible. There

are kits available for GC rich region PCR and sequencing and these may be required.

8.1.2.1. Alternative methods of replication

A discussion with Dr Tony Brooks at ICH, UCL Genomics led me to explore alternative methods for sequencing. The alternative methods were examined for not only coverage of the lead variant but also of the other lead 9 variants as these all had a SNV score of >200. Two different methods of hybrid capture sequencing were explored; NEB Next Direct predicted coverage of 7 out of 10 variants, whereas Nonacus estimated coverage of 2 out of the 10. The desired method for sequencing in a high GC content and repetitive element rich DNA segment is long read sequencing. Nanopore technology has the advantage of not requiring PCR amplification and single DNA molecule sequencing [389]. The long reads would improve the mapping and the insertions that are in the region [389]. Due to estimated costs these methods were not affordable.

8.1.2.2. Decision to stop further analyses

The difficulties encountered with alternative methods of replication and the poor coverage of the 10 lead variants with other standard techniques provided further evidence that the lead variant is not reliable and does not have a true association with AMN. There was a low SNV score and alternative sequencing methods were not able to provide good coverage. I therefore decided to stop pursuing analysis and replication of these results.

8.1.3. Limitations

There were limitations to the study due to the pooled DNA sequencing. The most significant and important limitation was the failure of one long range PCR reaction and the consequent loss of a 10,000 bp region near exon 1 (visualised in Figure 6.1). The first exon is particularly involved in expression of the gene; whereby the first exon length itself can contribute to the speed of protein expression [390]. Or the first exon's methylation status can affect translation [391]. Loss of the variants in this region is a great limitation and it is impossible to predict what valuable information was lost. Secondly, important haplotype frequency information was unavailable which may have helped determine the haploblock associated with AMN rather than a prediction based on control data [193]. Due to limitations with analysis and chi-squared testing, multi-allelic variants were excluded; potentially losing valuable information. The lack of PLA2R1 or THSD7A antibody status was another limitation; this may have strengthened the association with a *PLA2R1* variant if only the aPLA2Rab positive patients were investigated.

Another major limitation was that the control dataset was not sequenced on the same technology as our case dataset. This meant that valuable data was lost and the control data had to be filtered. The 1000 genome project used low coverage whole genome sequencing with imputation and therefore is not good quality deep coverage sequencing. There is mapping bias in the 1000 genome project data [392]. In the *PLA2R1* region, the 1000 genome project data has a greater discrepancy in incorrectly mapped variants of 5.15% compared to the whole chromosome of 2.9%. Incorrect mapping has been described in the HLA region and causes an error in the frequency estimation in the 1000 genome project dataset [392]. If these frequency

estimations are incorrect then we cannot rely on the difference to be indicative of a true variant in our MN dataset.

8.1.4. Re-analysing work

Given the opportunity to re-analyse the work now with the current knowledge I would make some changes to the workflow of analysis. Firstly, I would set an arbitrary scale for the SNV score to be greater to fulfil the criteria of being a more reliable 'real SNV'. I would also analyse only variants that are reported in dbSNP and exclude novel variants to avoid sequencing errors. Further exclusion of ambiguous SNVs would ensure the dataset was reliable and did not include SNVs incorrectly aligned to the reference genome by deCODE. This should facilitate replication. Ideally, if there was no limitation with expenditure I would re-sequence the entirety of the *PLA2R1* region with a more robust modern method of sequencing such as hybrid capture sequencing or nanopore sequencing.

A newer bioinformatic tool, called PAINTOR utilises the statistical data generated from an association study and integrates functional genomic data to prioritise variants for follow up analysis [393]. It then outputs a probability for a SNV to be causal for the trait of interest. This may be useful as an alternative method to analyse data for functional intronic variants with the existing data that is available without any extra expenditure. Due to time constraints in learning the methodology I was unable to undertake this.

8.1.5. Conclusion

A Sanger sequencing study of the *PLA2R1* coding region in 95 AMN patients failed to identify any single rare causative variant for AMN, instead identifying common variants that are found frequently in the healthy population. I replicated the findings of the two common coding variants in a larger cohort of 335 AMN individuals with both variants having a strong statistical association, p-value $<5 \times 10^{-8}$. I also identified an additional 107 intronic variants that were statistically associated with AMN.

The identified lead intronic variant is of uncertain reliability due to a low quality score and the difficulties encountered with replication. This suggests the pooled sequencing would have had similar difficulties and amplification bias. It is in a region that is difficult to sequence due to a high GC rich content and nearby short interspersed nuclear element. The lead variant if reliably implicated would have been interesting because it is located within a known DNA region identified by ENCODE ChIP-seq experiments as a TFBS for CEBPB. CEBPB is a transcription factor involved in a T-helper 2 cell response which is the main mechanism of autoimmunity in AMN.

For future studies *HLA* gene sequencing with *PLA2R1* sequencing would help identify potential shared risk genotypes in AMN. With the further work that I undertook I hypothesise that the lead *PLA2R1* variant is conditional on a specific HLA class II intra-locus configuration to cause disease. A combination of interactions may occur between *PLA2R1* and HLA risk variants and the environment to initiate AMN. It is possible that the common exonic variant when combined with the specific

HLA type is rare and this is what causes disease. While this data did not identify an intronic causative variant in AMN it successfully replicated the previously identified coding variants.

8.2. Association analysis

8.2.1. Genome wide association tests

This study replicates the findings of the two previous GWASs demonstrating that the two most strongly associated loci with AMN are *HLA-DQA1* and *PLA2R1* [73, 171].

8.2.1.1. *PLA2R1* locus association

In this study, the *PLA2R1* locus has two independent SNVs associated with AMN; rs3792189 and rs2292390. The association with PLA2R1 is strongly established with AMN and it was identified as the first podocytic antigen in AMN in adults [57]. The lead SNV is in the fourth intron of *PLA2R1*. The second SNV is in the third intron of *PLA2R1*. Both have similar signals with OR =1.72 and =1.31, respectively. It is the first time that two independent signals within *PLA2R1* have been identified with conditional analysis. This could be due to a higher proportion of anti-PLA2R antibody AMN cases within this dataset. Utilising the GRS work this is estimated to be over 72% of the total AMN cases. The additional risk loci in *PLA2R1* suggests that the pattern of association of this gene to disease may be more considerable and complex than initially envisioned.

8.2.1.2. HLA locus association

The strongest association with AMN remains in *HLA-DQA1* per this association study. The lead SNV is often in linkage disequilibrium with the true causative SNV and so as per the HLA association studies, this SNV is likely to be a marker for the two lead HLA types associated with AMN, see 8.2.2. The strength of the association remains very strong as previously identified by the two other GWASs [73, 171]. Even

with co-variate analysis with the 10 lead principal components the strength of the association remained high with an OR =2.6 in a heterozygous state. Conditional analysis demonstrated an additional SNV within chromosome 6 which was located near *HLA-DQB1*. Again, this is similar to the study by Xie *et al*., which reports 3 independent SNVs in the HLA region [171]. Reviewing the location of the Xie *et al.* SNVs in UCSC, the lead SNV is closest to *HLA-DQA1*, the second SNV to *HLA-B* and the third SNV to *HLA-DQB1*. These results are like the results presented in this study, although the *HLA-B* signal was not identified. This lends further support and evidence to the accuracy of these findings. The two independent SNVs in the HLA region support the two HLA types found in European AMN cases in this study. It is important to recall as previously discussed that even those these are being identified as independent associations these two HLA types form part of a tightly linked multigene haplotype, see 5.2.8.

My case cohort includes 111 non-European ancestry individuals, this is a mix of individuals with predominantly South Asian and African ancestry with a few admixed and East Asian ancestries, see Figure 7.29. Because of the small number of individuals in each of these non-European ancestries and more importantly the lack of reference controls for both imputation and for association testing, I was unable to analyse this cohort. It is likely that they have a different HLA type associated with AMN with some overlap to Europeans as Xie *et al.* found with the East Asian cohort. The difficulty with a smaller number of individuals is the lack of power to detect an association. However, ancestry differences can cause over- or underestimation of associations. This is because allele frequencies of variants differ across different ancestries and are not related to the phenotype itself [304]. In addition, different

ancestries are associated with different environmental exposures [394]. Further, while genetic population sub-structure can be accounted for with the statistical methods, it may not fully compensate for them [304]. Non-European, in particular African ancestries, have more genetic variation and this could help discover novel associations but the predominance of genetic analysis in Europeans hinders this discovery [395].

8.2.1.3. Unidentified associations

This study did not identify the two newest loci identified to be associated with AMN; *NFKB1* and *IRF4*. The reason for this is likely to be due an underpowered study as these two loci were only detected in the meta-analysis of both European and East Asian ancestry which totalled 3782 cases compared to 9038 controls. The European discovery-2 cohort had 1045 European cases and 1094 controls; similar to my case sample size. In the discovery cohort GWAS neither *NFKB1* nor *IRF4* reached statistical significance with a p-value of 0.05 and 0.02 respectively. *NFKB1* and *IRF4* were not identified in any of the sub-cohort analyses but only in the final meta-analysis of all ancestries. Another advantage was that the cases and controls were all genotyped on the same microarray chip which increased the pre-imputation intersecting SNVs which meant imputation was improved and the post-imputation genotyping data provided better resolution and coverage, see 8.2.4.2. Post imputation my analysis only had over 2 million SNVs whereas the Xie *et al.* study had over 6 million SNVs for analysis. I examined if any SNVs overlying these two genes were present in the association test results and there were none. Identifying an association would have been dependent on a few neighbouring gene linked SNVs which explains the difficulty in identifying the association. [171]

No further associations were found with clinical parameters in the 225 aPLA2Rab positive AMN cases. This could be due to a lack of power because of the low number of cases; which is a methodological problem. Or it may represent a weakness of association - i.e. there is no association between clinical parameters and aPLA2Rab AMN.

8.2.1.4. Antibody status

The genome wide association study with aPLA2Rab status strengthened the association with the two risk loci in *PLA2R1* and *HLA-DQA1*. The odds ratio for the risk of aPLA2Rab mediated disease increased from 3.29 in *HLA-DQA1* to 4.78. which is 1.45 times greater odds ratio than for all cases of AMN GWAS. The odds ratio with the *PLA2R1* variant increased from 1.72 to 2.03. This is striking and is a representation of the fact that the majority of cases in the unselected AMN case cohort are predominantly aPLA2Rab positive and drive this association. I hypothesise that these loci are specifically associated with aPLA2Rab positive AMN. This is corroborated by the differences seen in the different antibody status and GRS which also confirms that the aPLA2Rab positive group is genetically different to the other antibody groups, see 7.3.2. A suggestive locus in *TLR10* is present but does not reach genome wide statistical significance. This is 64 Mb away from *NFKB1* so is unlikely related to the previously identified chromosome 4 association [171].

Due to the low number of cases in the anti-THSD7A cohort, no statistically significant findings were made. This is due to the study being underpowered. Because it was difficult to obtain more cases (as it is a rare antibody) I decided to proceed with the

analysis with this in mind. The Michigan Genetic Association Study power calculator was used to calculate the number of cases required for a statistical power exceeding 0.8 [396, 397]. This calculated 160 anti-THSD7A antibody positive cases were necessary to detect a statistical difference of p-value $<5 \times 10^{-8}$. For this reason, meaningful conclusions cannot be drawn from this GWAS in the anti-THSD7A antibody positive group, which only contained 31 cases. Based on the differences seen in the GRS, I propose that the associations will be different to the aPLA2Rab group with a different HLA-type and potentially with *THSD7A* on chromosome 7.

Interestingly the association with the dual negative antibody cases simply identified the same lead loci as that of aPLA2Rab positive AMN. This association is most likely being driven by the proposed one third of dual negative cases that I hypothesise may be missed aPLA2Rab positive cases, see 8.3.1.2.

8.2.1.5. Clinical parameters

There were statistically significant single SNVs associated with uPCR or aPLA2Rab titres with p-values $=1.89 \times 10^{-8}$ and 3.99×10^{-8} respectively, see 7.2.6.5. Even though they reached statistical significance they are not true associations. In a GWAS you expect to see a build-up of SNVs in the peak because there are multiple intronic and exonic SNVs that are in linkage disequilibrium with the lead SNV. The closest gene located on chromosome 11 near the intronic SNV associated with uPCR is *LOC107984423* which is uncharacterised, so no information is known. The closest gene located on chromosome 4 downstream to the lead SNV associated with aPLA2Rab titres is the same gene, *LOC107984423*. The upstream gene is another uncharacterised gene called *LOC105369410*. As both genes are uncharacterised

and the association is weak, I do not think this is a true association but rather a false positive result.

8.2.2. HLA association tests

This study provides confirmation of the dominant HLA type in AMN. Statistically they were identified as two independent HLA types; *DRB1*03:01* and *DQA1*05:01*. These are the same two HLA types that had been identified in the most recent GWAS [171]. No further HLA associations were found using the T1DGC as the reference panel for imputation, see 8.2.4.4. I also demonstrate that these HLA associations are found in the aPLA2Rab AMN group. It is important to appreciate that these HLA types are part of the tightly linked common European multigene HLA haplotype and can not be accurately separated to determine which part of the haplotype confers disease risk [175-177]. This is a limitation of the statistical tests that are employed to analyse the genotyped and imputed data.

The HLA association test was underpowered to detect a difference in the HLA types in the anti-THSD7A antibody cases. The Michigan Genetic Association Study power calculator was again used for a statistical power exceeding 0.8 [396, 397]. This calculated that 90 anti-THSD7A antibody positive cases were necessary to detect a statistical association with a p-value <0.00043. For this reason, meaningful conclusions cannot be drawn about the HLA type of the anti-THSD7A antibody positive group, but I propose that this will be different to the aPLA2Rab group. Prior to correction for population stratification there was an association with the anti-THSD7A antibody AMN when HLA imputation was done with the T1DGC European reference panel. This is discussed further in 8.2.4.4. As this disappeared once

population stratification was accounted for, it represents a false positive due to population substructure differences.

8.2.3. Epistasis

Testing for pairwise epistasis was done with two different statistical models. Both confirmed an interactive effect between the *HLA-DQA1* and *PLA2R1* SNV with strong statistical significance. This phenomenon of epistasis in AMN has been described before and is not unexpected. In the most recent GWAS by Xie *et al.,* the epistatic interaction between the HLA risk haplotype and the *PLA2R1* risk allele was driven by the main HLA type of *HLA-DQB1* in Europeans. For double risk homozygotes the OR =14.1 [171], this is similar to the double risk homozygote OR calculated within this dataset =13.8. The phenomenon of epistasis has been described in smaller studies (280 Europeans) utilising a multifactor dimensionality reduction method for analysis which had a higher interaction OR =7.46 [92] (compared to 1.38 in my analysis). The strength of association in that smaller study may be higher because only aPLA2Rab positive AMN cases were analysed [92].

Epistasis testing only demonstrates a statistical interaction, but there is biological plausibility that a biological interaction could exist with these two loci. The presence of aPLA2Rab AMN is more likely in this study when both loci are present. The podocytic antigen PLA2R1 is different to healthy individuals however it is only when the common variant is present does it combine with the specific HLA-type that disease manifests. This is not surprising as HLA-DQA1 only recognises PLA2R1 when the variant is present and only then does it present PLA2R1 as an antigen to the immune system.

8.2.4. Limitations and challenges

8.2.4.1. SNV microarray and quality control

Quality control issues had been identified by colleagues working within our laboratory and this had resulted in the development of our in-house program Remedy to filter and exclude SNVs not passing quality control and not matching dbSNP. There was a high attrition rate following on from Remedy and after minor allele frequency filtering. This is a result of the SNV microarray chip design, which was designed specifically for multi-ethnic datasets and rare diseases and so contained multiple rare variants. If I were to repeat the genotyping I would select a different microarray SNV chip. A recent comparison of SNV microarrays can help select the one best suited for the study; in this instance I would prioritise a microarray designed not with the greater genome wide coverage but one designed with imputation quality in mind [398]. These issues were overcome by Xie *et al.* by genotyping controls on the same microarray chip so this would have been an alternative though costly mechanism by which to overcome this limitation. In retrospect, I also used a harsh MAF filter at the start of the QC process in the cases of 5%. This excluded a considerable proportion of SNVs which otherwise could have improved the overall number of intersecting variants between the control data prior to imputation.

8.2.4.2. Intersecting variants between case and control datasets

8.2.4.2.1. Pre-imputation

A considerable issue that resulted in the loss of a large number of SNVs was sourcing the control datasets from different publicly available databases. All three

control datasets and the cases were genotyped on different SNV microarray chips. As a result, despite genotyping 1,779,818 SNVs in the cases there were only 188,662 good quality SNVs that intersected with the control datasets, which means 89.4% of SNVs were lost from analysis. This problem was overcome by genome wide imputation of the combined case control dataset, which improved the density of SNV coverage. Despite this, the number of SNVs only increased to just over 2 million which is close to the starting genotyped number of SNVs. The same two loci (*HLA-DQA1* and *PLA2R1*) were identified in the unimputed dataset providing evidence that this was not just due to imputation error but are true associations.

8.2.4.2.2. Post-imputation

When the case and control datasets were imputed separately, there was considerable issues with population stratification. This is because they were genotyped on different microarray platforms and so genotyping differences already exist [304]. Combining these accentuates the differences and results in false-positive associations evidenced by multiple SNVs reaching statistical significance across the whole of chromosome 2 in my analysis, Figure 8.1. This was unexpected, as reports of successful separate imputation exist. One study reports low pre-imputation intersecting SNVs (approximately 38,000) with successful merging post imputation with the same (but less stringent) pre SNV QC steps [399]. The caveat to this is, they did not perform an association test and their post imputation principal component analysis demonstrated considerable divergence across all ancestries. Further resources on Biostars also suggest that it is possible to merge datasets after imputation [400]. I have not been able to find any published reports of post imputation dataset merging causing difficulties. However, at a genetics course, it was

stated that imputing cases and controls genotyped on different microarray chips separately induces spurious associations and considerable imputation artefacts [401]. The advice was to always merge the dataset if genotyped on different microarray platforms and impute together [401]. This is consistent with the subsequent results I obtained. This makes sense as imputation is used to amplify coverage and hence the signal. However, if spurious differences already exist it will also amplify these differences and result in noise.

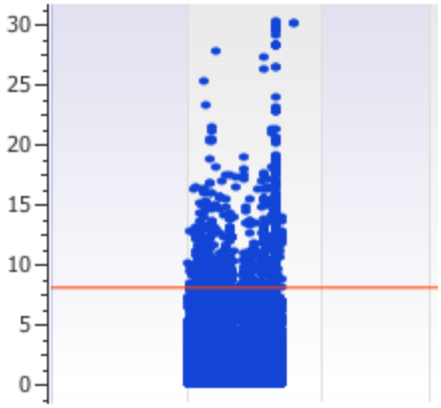


Figure 8.1: Manhattan plot of association test for chromosome 2 showing multiple false positives. Dataset comparing 1035 AMN cases versus 5080 controls imputed separately with a European reference ancestry panel and merged post-imputation.

8.2.4.3. Whole genome imputation reference panels

The 1000 Genome Project phase 3 reference panel for imputation is the most frequently used reference panel for whole genome imputation [309]. It provides data on 2504 individuals, however, only 503 of these individuals are European [309]. Because of the selection of predominantly European SNVs or non-ancestry specific SNVs in the 1000 Genome Project, it is a reference panel that can be used for European dataset imputation with accurate results [312]. For non-European datasets it is an inaccurate reference panel [312].

The decision between using the European reference panel or all ancestry non-matched reference panel for imputation is difficult. The European reference panel has fewer haplotypes, whereas the all ancestry reference panel has a larger sample size. Both a larger sample size and ancestry matching improve imputation accuracy; but with the available reference control panels they are achieved differently. Some studies report a higher success rate of good quality imputation with a better matched reference panel [312] and the European subset of the 1000 Genome Project has high genotyping rates and low imputation error rates [313]. Some studies chose to extract just the European individuals from the reference panel such as Xie *et al.* [171] whereas others utilise the whole dataset such as Dufek *et al.* [322]. For rarer variants, minor allele frequency <0.5%, imputation should only be done with ancestry specific reference panels [402, 403]. There is no consensus on the best method for imputation.

Because of the uncertainty in deciding between imputation with a smaller reference panel of ancestry matched or a larger reference panel of mixed ancestry I undertook imputation using both panels to compare the results for chromosome 2. There was a difference in the lead SNV identified with the different ancestry panels. Imputation with the European ancestry reference panel matched the expected signal within *PLA2R1*. The lead SNV in chromosome 2 imputed with all ancestry reference panel was rs4292050. This variant is intergenic, is 21.9 Kbp upstream of *PLA2R1* and is protective with an OR =0.55. Making the risk allele for AMN guanine which is the major allele (high population allele frequency of 70%) and an OR =1.82. This is unusual given the rarity of AMN as the theory of common variant common disease

has long held steadfast [404, 405]. To ensure this was not an error in my source code I rechecked all the scripts and datasets again. As a further check, I exported the dataset to run an association test in SVS which yielded similar results; lead SNV rs4292050, OR =0.51, p-value =2.54 $x10^{-9}$. It is interesting that there is a difference between the results for imputation between the two datasets and though they both implicate the same gene in the association study, the direction of association is different. This has not been described before but I postulate this may be because the risk variant could be linked to other protective variants in all ancestries and so the allele frequency remains high. In Europeans, the allele frequency of the risk variant ranges from 40-50% and so has become less frequent due to genetic drift as it is a smaller population. Further even though the risk variant is common it is the combination and interaction with the HLA-type that makes AMN a rare phenomenon.

8.2.4.4. HLA imputation reference panels

The HapMap is an older European HLA reference panel and the more recently available T1DGC is the other available European reference panel. The HapMap European reference panel has 124 individuals whereas the T1DGC has 5225 European individuals [315]. The recent GWAS by Xie *et al*. [171] used the T1DGC as their reference panel for HLA imputation. Using this reference panel for HLA imputation, it was possible to replicate the findings from this recent GWAS. The biggest concern with the T1DGC reference panel is that these individuals are disease free parents or siblings of an affected individual with type 1 diabetes mellitus. Due to the familial aggregation of class II HLA genes in type 1 diabetes this may not be the most suitable reference panel, but it is the most widely used and accepted [316]. The HapMap reference panel imputed to 56 4-digit types whereas

the T1DGC reference panel imputed to 114. Further, the association test results with the T1DGC reference panel matched those previously described, identifying the two lead HLA types; DQA1*0501 and DRB1*0301. The HapMap reference panel identified two previously unidentified HLA types; DQB1*0503 and A*0301. The HLA type of AMN had been demonstrated in previous studies as DQA1*0501 and DRB1*0301 [171], so this is surprising and suggests that the HapMap reference panel is not accurate for HLA imputation. For this reason, the imperfections with the T1DGC reference panel must be accepted.

8.2.4.5. Antibody status

Another limitation was the lack of detailed phenotype information; in particular the antibody status. Due to the historical multi-centre nature of the cohort, serum was unavailable in 279 individuals. Further the subdivision of the dual negative group was based exclusively on a single serum result. This is likely to have missed some cases of aPLA2Rab positive cases as it is recognised that a proportion of cases will only have renal biopsy staining for PLA2R. In active disease this difference is estimated at 4%, but in remission this is increased to 37% [75]. The German collaborators examined biopsies in a subset of 20 cases and confirmed that in these PLA2R was negative. Confirming in this subset that cases labelled as dual antibody negative were correctly identified. It was not possible to review the remainder of the cohort due to the different sources of DNA and lack of biopsy material available. This is a limitation as clear and accurate phenotype labelling are essential to determine genetic differences and would have strengthened the ability to detect associations.

8.2.5. Conclusion

This study confirms and replicates the findings that AMN is associated with *PLA2R1*, *HLA-DQA1* and *HLA-DRB1* in a European cohort. In particular, I establish for the first time the HLA type associated with aPLA2Rab AMN as *DRB1*03:01* and *DQA1*05:*01. Both HLA types are not predominant in anti-THSD7A antibody AMN however because this analysis was underpowered meaningful conclusions cannot be drawn. The alleles in both loci are common, however when they combine together they are rare and therefore increase the preponderance to a rare disease: AMN. There is a demonstrable statistical interaction between *HLA-DQA1* and *PLA2R1* in epistatic testing. While the variants are likely not directly causative, the presence of both risk alleles in both loci increases the risk of AMN by 14-fold.

8.3. Genetic risk score

8.3.1. Antibodies & genetic risk

8.3.1.1. PLA2R and THSD7A antibody mediated AMN

This study confirms there are genetic differences between different autoantigen-defined subgroups of autoimmune MN. Despite sharing a clinical and pathological phenotype, the aPLA2Rab group is genetically distinguishable from the anti-THSD7A group. It may be expected that the two antibody groups would differ most significantly at the *PLA2R1* locus because it has been suggested that the variation in the *PLA2R1* sequence alters PLA2R1 expression and increases the risk of being presented as an autoantigen [72]. Missense SNVs within the C-type lectin domain in *PLA2R1* are predicted to affect the binding affinity of *DRB1*1501* [166] and create a T-cell epitope for presentation to *DRB1*1501* in a Japanese cohort [95]. Interestingly, the difference in *PLA2R1* allele count between the anti-PLA2R and anti-THSD7A antibody groups did not reach statistical significance. Instead, variation at the *HLA-DQA1* locus was more strongly associated with antibody status. The risk effect of HLA for aPLA2Rab AMN could be explained by HLA antigen specificity, whereby the associated allomorph is more likely than others to present self-PLA2R1 peptide(s), but not more likely to present self-THSD7A epitopes. HLA peptide recognition is known to be highly specific in other kidney diseases [406] so a similar method may be present in aPLA2Rab AMN. In ANCA-associated vasculitis the different specificity of autoantibodies is associated with distinct subsets of HLA class II allomorphs [407].

8.3.1.2. Dual antibody negative AMN

A proportion of the dual antibody negative group have a similar GRS to that of the aPLA2Rab group. This contrasts with previous work that only found the *PLA2R1* risk allele association in aPLA2Rab positive cases [92]; although PLA2R1 positivity was detected with biopsy immunofluorescence in that study. This may be explained by imperfect sensitivity of the serum assay to detect anti-PLA2R antibodies in some patients, loss of antibody due to prolonged storage of serum samples or to patients becoming seronegative between onset of disease and sampling of their serum [162]. Some antibody-negative cases might be associated with the rarer antigens for which we did not test, such as NEP, HTRA1, PCDH7, Sema3B, EXT and NELL-1. Nonetheless, the elevation in AMN GRS among this antibody-negative group suggests that a significant proportion of the dual antibody negative patients have disease driven by PLA2R1 immunoreactivity. This is supported by previous data showing that 11% of individuals serologically negative for aPLA2Rab show immunoreactivity histologically [408, 409], although in my cohort of dual antibody negative cases (N =384) this proportion on its own would not completely account for the similarity in GRS I observed between aPLA2Rab positive and antibody negative groups.

The sensitivity of serological testing is variable and previous estimates range between 50-80% [410]. Simulation analyses (comparing randomly drawn GRS from samples of aPLA2Rab positive and healthy control subjects in different proportions) indicated that the GRS observed in the antibody negative group is best explained by this group comprising approximately 33% aPLA2Rab positive individuals. Since I did not observe a correlation between aPLA2Rab titre and GRS I regard this as the best

estimate for the proportion of apparently antibody negative individuals who actually have disease driven by autoimmunity to PLA2R1. This equates to 127 individuals and suggests (assuming a specificity of ≥97% as previously reported [411]) that the sensitivity of the PLA2R1 assay used to detect PLA2R1 autoimmunity was 76% (127 out of 533 proven and inferred PLA2R1 positive cases not detected). Our collaborators in Germany reviewed a small proportion of the biopsies available from the dual negative cases and confirmed that the biopsy PLA2R1 immunofluorescence was negative [412]. This corroborates recent findings suggesting the GRS can help establish a diagnosis in dual negative MN in 20-37% of cases [171].

8.3.1.3. Anti-contactin antibody associated AMN

In a small cohort of 6 European individuals with anti-contactin associated AMN and CIDP I was able to identify that the genetic risk is different to aPLA2Rab associated AMN. This provides further evidence to the specificity of the GRS for aPLA2Rab mediated disease. Interestingly, and different to anti-THSD7A AMN, anti-contactin AMN was not statistically different in the *HLA-DQA1* risk variant but instead in the *PLA2R1* risk variant.

The contrasting differences found between the anti-THSD7A and anti-contactin genetic associations are difficult to understand. It is likely that the results represent either a false negative or a false positive because of the low number of anti-contactin cases; so, meaningful and reliable conclusions cannot be drawn. Yet, if the results are reliable, then it suggests that the dominant HLA type in anti-contactin cases may be the same as in aPLA2Rab AMN, although it should be noted that there were no homozygotes for *HLA-DQA1* in the anti-contactin cohort. It would be interesting to

undertake HLA association in these anti-contactin cases; this was not done currently as the analysis would be underpowered. Alternatively, this finding refutes my prior hypothesis that HLA specificity is involved exclusively in the presentation of PLA2R1. The genetic locations of all three antigens investigated here differ with *PLA2R1* on chromosome 2, *THSD7A* on chromosome 7 and *CNTN1* on chromosome 12 [413, 414]. At the structural level THSD7A and PLA2R1 have similar structural and biochemical properties [98, 415]. They both have high molecular masses (210 kDa and 180 kDa respectively) compared to contactin-1 which has a molecular weight of 113 kDa [413, 414]. The overall structure of contactin-1, however, appears very different to PLA2R1 and THSD7A [415]. It may be this difference in protein structure that explains the difference seen in the HLA specificity.

8.3.2. Age and HLA

The *HLA DQA1*05:01* risk allele is associated with a younger age of onset of disease in aPLA2Rab positive cases, with an inverse correlation between age and allele count observed in both European cohorts studied. A similar finding, but with a different HLA type (HLA *DRB1*15:01* positive and *DRB3*02:02* negative) was demonstrated in a Chinese cohort of 100 aPLA2Rab positive cases [158]. No association with GRS and age was found in AMN in 1752 unselected primary MN cases [171]. With the inclusion of only aPLA2Rab positive cases, this association was detected in this study. The phenomenon of HLA risk alleles being associated with age of onset has been described in paediatric cases of SSNS where increased number of risk alleles in *HLA-DRB1* and *DQB1* was associated with a younger age of disease onset [416].

In non-nephrotic diseases, this phenomenon has also been described, for example homozygosity at three risk HLA alleles (two class II and one class I) was associated with younger age of onset of diabetic related ESKD in Saskatchewans (Indigenous Canadians) [417]. The increased genetic risk increases the likelihood that the sum of risk factors exceeds the threshold for disease manifestations: the disease is more likely to happen and thus tends to occur earlier in life. One review speculates (without any evidence) that risk genes cause genetic anticipation of AMN such that in subsequent generations it occurs at a younger age [418]. I do not think this is accurate as there is no evidence that the median age for AMN has changed over the past few years, and in this study the median age at onset of disease is 57.

8.3.3. Paediatric onset AMN

In 15 cases of non-familial aPLA2Rab negative paediatric onset AMN the GRS was different to aPLA2Rab AMN. These individuals were also negative for THSD7A but the status of other rarer antigens was not known. Of particular interest and reported after this work was undertaken, was the finding of Sema3B, see 5.1.7.5. This work identifies the strong association of Sema3B with paediatric onset AMN and so it is possible that my cohort may have antibodies against Sema3B. It would be ideal if saved serum is available to measure Sema3B levels.

The paediatric cohort was different in the GRS, the *PLA2R1* allele count and the *HLA-DQA1* allele count, confirming in all three components their genetic difference to the aPLA2Rab group. Calculating the SSNS GRS provided evidence that these paediatric onset AMN cases were not misdiagnosed SSNS cases.

My study contrasts with that of a larger cohort of 18 paediatric onset AMN cases (age range 10-19 years) [419]. In that cohort aPLA2Rab was detectable in 83% and PLA2R1 was visible on renal biopsy in 77.8% of cases [419]. My paediatric cohort were recruited as paediatric cases of non-familial AMN. After recruitment serological testing confirmed their dual antibody negative status. An explanation for this difference could be that the collaborators included the aPLA2Rab paediatric onset cases within the adult aPLA2Rab cohort. Another explanation is that in the study by Kumar *et al.* most cases were adolescent onset AMN with a mean age of 16 so they are more likely to have a greater proportion of aPLA2Rab associated disease because they are more similar to adults.

8.3.4. Use of GRS

It is desirable to assess the contribution of genetic risk to disease risk as this is directly and easily measurable; the method proven to have clinical utility is the GRS, see 5.4.2.4.

8.3.4.1. Interpreting a GRS score

The GRS has key components to consider for its interpretation. At a population level it can be informative to determine the risk of a certain phenotype. In this study, the GRS predicts the risk of aPLA2Rab associated AMN. This is useful for informative purposes but identification of individuals that are a greater risk for AMN will not results in specific screening or preventative therapeutic or behavioural intervention. The onset of AMN is unpredictable and it is possible that, despite having the genetic risk factors for aPLA2Rab AMN, an individual themselves will never be afflicted with disease. The GRS may be useful as a comparator for an individual but cannot

accurately predict true risk. An online tool, called the Polygenic Score Catalog, curates information on diseases and associated GRS scores and the SNVs and odds ratio are available to be downloaded [420]. As an example, a recent entry for the GRS in chronic kidney disease highlights 183,272 SNVs used to calculate the genetic risk [420].

8.3.4.2. Utility of GRS at an individual level

Despite strong evidence and association of disease status and GRS it is difficult at an individual level to state what the risk of disease would be to that individual. A suggested way to facilitate the conversion of the GRS to a clinical tool is to calculate relative and absolute risks which would be more informative at an individual level [233]. Another way is to use the GRS with other risk prediction tools and create it as an additional feature as part of that tool [233]. This is not possible in AMN as no risk prediction tools exist for onset of disease, so it could not be easily utilised as in other diseases such as coronary artery disease. Xie *et al.* develop a combined risk score (CRS) which combines the GRS and the serum aPLA2Rab titres [171]. The authors state that by using the CRS the need for a biopsy could be obviated [171]. However, the added advantage to the CRS compared to aPLA2Rab serum titres is minimal. The specificity for aPLA2Rab titre measurement was 100% and for the CRS 99%. Further the sensitivity of aPLA2Rab titre measurement was 57% which increased to 60% with the CRS [171]. No sensitivity or specificity data is provided for the GRS they developed nor comparisons of this compared to aPLA2Rab serum titres [171]. The area under the receiver operating characteristics curve (AUROC) for GRS in Europeans is 0.75 [171] and a meta-analysis predicted the AUROC for aPLA2Rab titres at 0.82 [408]. This suggests that the GRS is similar to aPLA2Rab titres in

correctly predicting an AMN case. There may be a use for the GRS to the individual but with the information that can currently be gleaned with aPLA2Rab titres it is more beneficial in identifying the incorrectly diagnosed dual antibody negative AMN cases. At present, in AMN, providing individuals with their GRS may not prove beneficial for them and instead may bring anxiety and concern about genetic determinism and so the sharing of individual risk would need to be done with specific counselling to prevent this from happening.

Another way to utilise the GRS at an individual level would be to undertake testing or reporting test results at an appropriate time triggered by onset of symptoms, family history, age or environmental factors [227]. In AMN it could be used in cases of aPLA2Rab negativity. If the genetic risk is high despite negative serology, it suggests they have aPLA2Rab mediated disease and so should be treated if relevant with immunosuppression. This would prevent unnecessary delay in treatment and unnecessary investigations from radiation such as CT scanning that is done in cases of antibody negative AMN to exclude secondary causes.

The widespread use of GRS is normally for screening for primary prevention [233]. An example of this is in coronary artery disease. The highest GRS quintile group had a hazard ratio of 1.66 for a coronary event [421]. A primary prevention strategy with statin therapy led to a 45% reduction in the relative risk in the high risk group compared to 24% in the other lower risk GRS groups [421]. A primary prevention strategy is not relevant in AMN because no screening is routinely undertaken for AMN nor would it change disease onset or outcomes. The closest example of screening is in Japan where routine urine dipstick screening is done; however, they

find a lower proportion of aPLA2Rab AMN [162, 163]. It is unclear if this is because of screening but rates of nephrotic range proteinuria are lower than other studies suggesting that patients are identified with milder disease and so potentially before aPLA2Rab are detectable [422].

8.3.4.3. Commercial use of GRS

GRS and ease of obtaining genetic data has gained considerable interest over the past few years with direct-to-consumer testing and marketing. With the rise of commercial platforms such as 23andMe, Nebula Genomics and others that provide DNA testing at home, access to genetic data at an individual level is increasing. Some of these commercial companies report GRS in diseases such as type 2 diabetes mellitus, Crohn's disease and atrial fibrillation [423]. This has been particularly contested because these companies create their own risk score from publicly available data from GWAS results from as few as a dozen SNVs for the risk of each disease [423]. The overall number of SNVs used for the GRS is not a concern, however, an attempt to replicate the GRS algorithms demonstrated large variability in the predictive ability for these scores [423]. This is very risky and has implications in inaccurate results being provided to an individual.

8.3.4.4. Environment and GRS

The GRS does not take in to account potential gene and environment interactions. Having a genetic risk does not mean that disease will ensue and certainly contributory environmental exposures are necessary.

The role for environmental factors in AMN is not fully elucidated, but there is predicted to be an environmental trigger for disease. In China, the increasing prevalence of AMN was associated with an increase in fine particulate matter ($PM_{2.5}$) pollution [152]. Another study identified AMN patients had a worse prognosis if they had higher levels of Th17 mediated inflammation, which was related to higher levels of pollution [424]. Inflammatory triggers are thought to have a contributory role and the overlap of inflammatory bowel disease and the AMN genetic risk loci suggests a shared mechanism [171]. It is hypothesised that *PLA2R1* is upregulated in an inflammatory environment and self-tolerance abilities are lost [425]. Couser and Johnson [426] hypothesise that firstly, genetic risk factors increase predisposition to disease and the response to an environmental factor. Secondly, the exposure to the environmental factor activates the immune system via separate epigenetic factors [426]. Both then combine to cause autoimmune kidney disease such as AMN.

8.3.5. Further work

To further develop this aspect of the study it would be useful to calculate the GRS in other ancestries. Within the European ancestry there can be a spread in genetic diversity and so the target population may not overlap well with the source GWAS population. This phenomenon was reported when a European GRS was examined in UK, Estonian and German European populations and it was noted that there was a difference in the estimation models and applicability because of population differences, despite European ancestry matching [427]. To overcome and investigate this further, it would be interesting to ancestry match to the original GWAS dataset from 2011.

GWAS studies in non-European AMN are limited, but most recently the identification of four lead SNVs in an East Asian AMN population was made [171]. A particular difficulty with this is that the cohort used in this current study has few East Asian individuals (based on observations as seen in Figure 7.29), so this would be difficult to undertake within this current cohort. One of the strengths of this cohort, is that there is a large proportion of South Asians (visualised by the tail extending upwards towards the East Asians in Figure 7.29). A GWAS would first need to be conducted and then a GRS model could be proposed and validated in a separate cohort. The same could also be done for African AMN individuals. It would be interesting to see if these individuals had any additional genetic risk factors that increase the chances of reaching ESKD.

It would be interesting to apply the GRS with the recently identified 5 lead AMN SNVs [171]. This was not done because the genotyping data from my cohort was used in this GWAS. This would have meant that the target population and the source population for the GRS analysis is the same and would naturally be very biased. To overcome this, an alternative methodology suggests adjusting the GWAS odds ratio to mitigate the 'winner's curse' [428]. The authors suggest three methods to reduce bias estimators, and the one that is the most unbiased is a method called the 'mean square error'. This is a weighted average of the corrected and uncorrected odds ratio and prevents both over and under correction [428].

The paediatric cohort is unique and is the largest cohort reported of dual antibody negative paediatric onset AMN. The overall number will be too low to detect a statistical association(s) to reach genome wide significance but it would be

interesting nonetheless to undertake a case-control GWAS to see if this identifies any further associations and/or confirms the association with Sema3B. The HLA risk SNV of this cohort is different to the AMN risk SNV; investigating this by first HLA imputation and then association testing would be interesting to determine if a predominant HLA type was demonstrable. The same could be done with the anti-contactin antibody group.

8.3.6. Limitations

The greatest limitation with the study was the lack of detailed phenotype information across the dataset. Due to the historical multi-centre nature of the cohort, serum was unavailable in 279 individuals. This would have helped improve the overall numbers and increase the statistical power to detect an association. Other limitations were that in 97 individuals genotyping data was not available in the two loci and so they had to be excluded from analysis.

8.3.7. Conclusion

I observed genetic differences between aPLA2Rab and anti-THSD7A antibody positive AMN in European populations. Additionally, the similarity of GRS between aPLA2Rab and dual antibody negative AMN cases suggests that approximately a third of the antibody negative cases represent false negative serology. This implies that a negative aPLA2Rab assay alone should not be used to determine diagnosis or treatment. With this work I also demonstrate that application of the GRS in AMN can distinguish different antibody states even in the presence of false negative serological testing. The clinical utility of genetic risk scores is established but further

clinical studies are needed to determine whether a GRS in AMN can provide clinical benefit to individual patients.

8.4. UK Biobank

In this study the UKBB dataset is analysed for the European AMN GRS in 419,802 individuals. This demonstrates that 0.8% (3,372) of European individuals in the UKBB have a higher genetic risk for developing aPLA2Rab associated AMN. The *PLA2R1* risk SNV is common in European individuals and is present in 33.9% (142,415 individuals) in homozygous state, whereas the *HLA-DQA1* risk SNV is rarer and is present in only 2.39% (10,019). The number of individuals at risk of AMN is high, and higher than the estimate with the current incidence of 10 to 12 per million per population. Accurate prevalence data for AMN does not exist and data is only available in individuals with ESKD. With this current analysis and the current population of 66.65 million people in the UK; 533,200 individuals have a high genetic risk of developing AMN. Extrapolating this population genetic risk to the number of patients that would eventually reach ESKD would mean 159,960 individuals on dialysis in the UK. This would be a huge burden as the current number of total dialysis patients is approximately 30,000 for all diseases in the UK. Mitigating the genetic risk would be difficult and predictive and prevention tools for AMN are difficult. It is important to determine the environmental factors that are contributory to AMN as discussed in 8.3.4.4.

Genetic risk does not equate to certainty of disease onset, however, I would have expected this to be lower and more in keeping with the rarity of the disease. Arguably, this finding of a high population genetic risk is not unexpected and it has been demonstrated in type 1 diabetes mellitus. The GRS is utilised in type 1 diabetes; individuals with a high score (>99.9th centile) have a higher genetic risk for

disease onset (>20%) but this only identifies 7% of subsequent type 1 diabetes cases [429]. This is similar to the discrepancy that is seen in this AMN analysis. Rose [430] describes issues related to individuals in the population with a high risk for diseases. He explains because the number of individuals with minimal risk are more numerous, the numbers of those individuals with disease will actually be higher [430]. To summarise, ‘a large number of people at small risk may give rise to more cases of disease than the small number who are at high risk.’ [430]

The utility of GRS is discussed above, see 8.3.4. On a population level, identification of individuals at high genetic risk for AMN would be useful if there was a preventative or screening option for individuals that would minimise the risk of disease. For screening purposes, it may be possible to provide urine dipstick testing for individuals at greater genetic risk. This would cause a lot of over testing however and treatment strategies are unlikely to change despite seeking earlier medical intervention. With other GRS the preventative intervention is considered relatively innocuous (for example taking aspirin to lower the risk of coronary artery disease) and therefore is more amenable to widespread use at a population level. The main utility would be the overall cost savings that could be made in dual antibody negative cases of AMN. It has been proven that correct confirmation of aPLA2Rab associated AMN has a significant cost saving as it prevents a large array of unnecessary tests and screening for secondary causes [431].

8.4.1. Limitations

8.4.1.1. Population stratification

The GRS score is only valid in those of European ancestry and so it was desirable to extract only those individuals from the UKBB. There were considerable issues with this due to the size of the dataset and the frequently used tools being unsuitable. GRS is only applicable in ancestry matched populations; self-reported ancestry is not accurate so it was desirable to undertake this with a PCA method. Pre-computed principal components and ancestry cut-offs were not made available [352]. SmartPCA predicted 51 years for calculations with the dataset, so was unusable [334]. The downloadable code for bignspr did not work and again data used for the exclusion of outliers was not made available [353]. There are newer non-PCA methods for ancestry discovery and these could be explored, see 8.4.2.1.

UKBB state the self-reported ancestry of 'White British' has similar genetic ancestry on PCA. Due to the unsuccessful attempts of multiple resources in the interests of time I proceeded to use self-reported ancestry for analysis. This is frequently done by other published GWAS studies.

Analysing the spread of ancestry from the principal components, Figure 7.60, the split of the 1000 Genomes Project European reference control is evident. This is most likely because of the diversity across the European continent and in particular differences that occur between North and South Europeans [432].

8.4.1.2. AMN patients in UKBB

A limitation within the UKBB is that there are only 29 AMN patients within the dataset. This could be due to missing or incorrect hospital episode statistics not capturing the correct code for AMN. This number is lower than expected but it would be interesting to examine the AMN GRS in these individuals. Following on with further work after the completion of my thesis I was able to further investigate and identify more individuals with AMN, see 7.4.3.

8.4.1.3. Imputed dataset size

The imputed dataset when uncompressed was at least 12 terabytes so could not be loaded; this rendered the data inaccessible for use. This could be overcome with specialist expertise which would expand the use for the UKBB.

8.4.2. Further work

8.4.2.1. Non-PCA techniques of ancestry

It is desirable to have a more robust method of ancestry identification for the UKBB. This would facilitate accurate European ancestry identification and creation of a control cohort for South Asians and African ancestries for the GWAS studies and HLA association studies.

Newer techniques available that use a non-PCA method for ancestry discovery are iAdmix and GRAF-pop. iAdmix infers the proportion of admixture per individual using a likelihood ratio method based on a reference set of population allele frequencies. It takes 5.2 seconds per sample for analysis which for the UKBB dataset would take

approximately 30 days [433]. GRAF-pop uses a geometric distance-based approach to predict ancestry, accurately identifying the three continental populations of European, African and Asian. It is a useful tool but is limited as it is slower taking 75 seconds per sample in addition to being unable to accurately delineate South Asians or admixed individuals [434]. As a result it would not be suitable for use in the UKBB dataset.

8.4.2.2. Data with improved genome coverage

Analysis using the imputed dataset would increase the coverage across the genome and the intersection of SNVs with the AMN dataset. However, based on the findings in my work the differences would be amplified and so imputation would need to be done on the combined dataset. Use of the UKBB whole genome sequenced data would overcome this so that all SNVs in the AMN imputed dataset could be extracted from the UKBB and the rare variants could also be examined.

8.4.3. Conclusion

The UKBB dataset was used as a representative population for the United Kingdom. In individuals from a self-reported European ancestry, I observed that the genetic risk of developing aPLA2Rab AMN is 0.8%. Interestingly, despite this population's higher genetic risk, only 29 individuals had AMN which is 0.86% of all those at high genetic risk. This could be due to missed cases of AMN, because AMN is a later adult-onset disease so disease had not yet developed or the environmental factors that contribute to disease. This further demonstrates the absence of genetic determinism and the role for future studies to delineate contributory factors for disease onset.

8.5. Summary

This thesis has undertaken a multi-faceted approach to delineate the genetic contributors to autoimmune membranous nephropathy. I was able to replicate the findings of common coding variants in *PLA2R1* in disease. Rarer intronic variants were identified but it was beyond the scope of my study to investigate these all. The improving availability of cohorts with whole genome sequencing will facilitate repeat analysis and re-sequencing of *PLA2R1* to identify disease contributory variants. Because of the lack of coding variants, I hypothesise that the variant(s) will be involved in the regulatory pathways and disruption of these contributes to disease.

I was also able to replicate the GWAS findings from 2011 and partially replicate the more recent GWAS findings from 2020. I demonstrate in a cohort of European individuals that the strongest association with disease is with the *HLA-DQA1, HLA-DRB1* HLA haplotype and *PLA2R1*. I hypothesise that the GRS is specific for identification of aPLA2Rab associated AMN and can be used in serologically dual antibody negative patients to identify missed aPLA2Rab associated AMN cases and prevent unnecessary costly and radiating investigations. *HLA-DQA1* risk variants are associated with a younger age of onset of disease in aPLA2Rab AMN. Dual antibody negative paediatric individuals have a different genetic risk score to adult onset aPLA2Rab AMN and may represent Sema3B associated AMN cases. Anti-contactin antibody AMN cases also have a different genetic risk for AMN compared to aPLA2Rab cases. Finally, in the UK Biobank the proportion of individuals with a high genetic risk for developing AMN is 0.8% although of these only 0.86% have disease

at present. It will be interesting to investigate determinants of disease onset so that useful predictive and preventative strategies could be instigated.

9. Bibliography

1. McGrogan, A., C.F. Franssen, and C.S. de Vries, *The incidence of primary glomerulonephritis worldwide: a systematic review of the literature.* Nephrol Dial Transplant, 2011. **26**(2): p. 414-30.
2. Couser, W.G., *Primary Membranous Nephropathy.* Clinical Journal of the American Society of Nephrology, 2017. **12**(6): p. 983-997.
3. Lai, W.L., et al., *Membranous nephropathy: A review on the pathogenesis, diagnosis, and treatment.* Journal of the Formosan Medical Association, 2015. **114**(2): p. 102-111.
4. Jha, V., et al., *A randomized, controlled trial of steroids and cyclophosphamide in adults with nephrotic syndrome caused by idiopathic membranous nephropathy.* J Am Soc Nephrol, 2007. **18**(6): p. 1899-904.
5. Hogan, S.L., et al., *A review of therapeutic studies of idiopathic membranous glomerulopathy.* Am J Kidney Dis, 1995. **25**(6): p. 862-75.
6. Bradley, S.E. and C.J. Tyson, *The "Nephrotic Syndrome".* New England Journal of Medicine, 1948. **238**(7): p. 223-227.
7. Maisonneuve, P., et al., *Distribution of primary renal diseases leading to end-stage renal failure in the United States, Europe, and Australia/New Zealand: Results from an international comparative study.* American Journal of Kidney Diseases, 2000. **35**(1): p. 157-165.
8. Xu, X., et al., *Long-Term Exposure to Air Pollution and Increased Risk of Membranous Nephropathy in China.* Journal of the American Society of Nephrology, 2016. **27**(12): p. 3739-3746.
9. Mathieson, P.W., *Membranous Nephropathy*, in *Practical Nephrology*, M. Harber, Editor. 2014, Springer, London. p. XIX, 901.
10. Ponticelli, C., R.J. Glassock, and P. Passerini, *Membranous Nephropathy*, in *Treatment of Primary Glomerulonephritis*, C. Ponticelli and R.J. Glassock, Editors. 2019, Oxford University Press.
11. Glassock, R.J., *Diagnosis and natural course of membranous nephropathy.* Seminars in Nephrology, 2003. **23**(4): p. 324-332.
12. Gupta, S., et al., *Genetics of membranous nephropathy.* Nephrology Dialysis Transplantation, 2017. **33**(9): p. 1493-1502.
13. Lockshin, M.D., *Sex Differences in Autoimmune Disease*, in *Handbook of Systemic Autoimmune Diseases*, M. Lockshin, D.W. Branch, and R.A. Asherson, Editors. 2005, Elsevier. p. 3-10.
14. Alsharhan, L. and L.H. Beck, *Membranous Nephropathy: Core Curriculum 2021.* American Journal of Kidney Diseases, 2021. **77**(3): p. 440-453.
15. Moroni, G. and C. Ponticelli, *Secondary Membranous Nephropathy. A Narrative Review.* Frontiers in Medicine, 2020. **7**(928).
16. Leeaphorn, N., et al., *Prevalence of cancer in membranous nephropathy: a systematic review and meta-analysis of observational studies.* Am J Nephrol, 2014. **40**(1): p. 29-35.
17. Bjorneklett, R., et al., *Long-term risk of cancer in membranous nephropathy patients.* Am J Kidney Dis, 2007. **50**(3): p. 396-403.
18. Beck, L.H., *Membranous Nephropathy and Malignancy.* Seminars in Nephrology, 2010. **30**(6): p. 635-644.
19. Ion, O., et al., *Kidney Involvement in Hypocomplementemic Urticarial Vasculitis Syndrome-A Case-Based Review.* J Clin Med, 2020. **9**(7).

20. Aydi, Z., et al., *[Systemic sarcoidosis and membranous glomerulonephritis].* Rev Pneumol Clin, 2014. **70**(6): p. 375-9.
21. Bhimma, R. and H.M. Coovadia, *Hepatitis B Virus-Associated Nephropathy.* American Journal of Nephrology, 2004. **24**(2): p. 198-211.
22. Yang, Y., et al., *A Meta-Analysis of Antiviral Therapy for Hepatitis B Virus-Associated Membranous Nephropathy.* PLoS One, 2016. **11**(9): p. e0160437.
23. van Velthuysen, M.L. and S. Florquin, *Glomerulopathy associated with parasitic infections.* Clin Microbiol Rev, 2000. **13**(1): p. 55-66, table of contents.
24. Glassock, R.J., *Secondary membranous glomerulonephritis.* Nephrol Dial Transplant, 1992. **7 Suppl 1**: p. 64-71.
25. Rihova, Z., et al., *Secondary membranous nephropathy--one center experience.* Ren Fail, 2005. **27**(4): p. 397-402.
26. Hogan, J.J., G.S. Markowitz, and J. Radhakrishnan, *Drug-induced glomerular disease: immune-mediated injury.* Clin J Am Soc Nephrol, 2015. **10**(7): p. 1300-10.
27. Radford, M.G., Jr., et al., *Reversible membranous nephropathy associated with the use of nonsteroidal anti-inflammatory drugs.* Jama, 1996. **276**(6): p. 466-9.
28. Hoorntje, S.J., et al., *Immune-complex glomerulopathy in patients treated with captopril.* Lancet, 1980. **1**(8180): p. 1212-5.
29. Debiec, H., et al., *Antenatal Membranous Glomerulonephritis Due to Anti–Neutral Endopeptidase Antibodies.* New England Journal of Medicine, 2002. **346**(26): p. 2053-2060.
30. Honda, K., et al., *De novo membranous nephropathy and antibody-mediated rejection in transplanted kidney.* Clinical Transplantation, 2011. **25**(2): p. 191-200.
31. Barbara, J.A., et al., *Membranous nephropathy with graft-versus-host disease in a bone marrow transplant recipient.* Clinical nephrology, 1992. **37**(3): p. 115-118.
32. Boss, J., *Richard Bright's Reports of Medical Cases (1827): a sesquicentennial note.* Bristol medico-chirurgical journal (1963), 1978. **93**(345-346): p. 5-18.
33. Peitzman, S.J., *From Dropsy to Bright's Disease to End-Stage Renal Disease.* The Milbank Quarterly, 1989. **67**: p. 16-32.
34. Peitzman, S.J., *Dropsy, Dialysis, Transplant: A Short History of Failing Kidneys*. 2007: JHU Press. 213.
35. Cullen, W., *First lines of the practice of physic*. 3 ed. Vol. 1. 1784, Edinburgh: C. Elliot.
36. Hippocrates, J. Chadwick, and W. Mann, *The Medical works of Hippocrates : a new translation from the original Greek made especially for English readers* . 1950: Blackwell, Oxford.
37. Dekkers, F., *Exercitationes practicae circa Medendi Methodum.* 1694: Boutensteyn and Luchtmans, Leiden.
38. Cameron, J.S., *The History of Proteinuria.* Proteins in Normal and Pathological Urine, 1970: p. 1-5.
39. Cameron, J.S., *Milk or albumin? The history of proteinuria before Richard Bright.* Nephrology Dialysis Transplantation, 2003. **18**(7): p. 1281-1285.
40. Kark, R.M., *A prospect of Richard Bright on the centenary of his death, December 16, 1958.* Am J Med, 1958. **25**(6): p. 819-24.

41. Glassock, R.J., *The Pathogenesis of Idiopathic Membranous Nephropathy: A 50-Year Odyssey.* American Journal of Kidney Diseases, 2010. **56**(1): p. 157-167.
42. Jones, D.B., *Nephrotic glomerulonephritis.* The American journal of pathology, 1957. **33**(2): p. 313-329.
43. Farrow, B.R. and C.R. Huxtable, *Membranous nephropathy and the nephrotic syndrome in the cat.* J Comp Pathol, 1971. **81**(4): p. 463-7.
44. Orth, S.R. and E. Ritz, *The Nephrotic Syndrome.* New England Journal of Medicine, 1998. **338**(17): p. 1202-1211.
45. Doucet, A., G. Favre, and G. Deschênes, *Molecular mechanism of edema formation in nephrotic syndrome: therapeutic implications.* Pediatr Nephrol, 2007. **22**(12): p. 1983-90.
46. McCloskey, O. and A.P. Maxwell, *Diagnosis and management of nephrotic syndrome.* Practitioner, 2017. **261**(1801): p. 11-5.
47. Glassock, R.J., et al., *Nephrotic syndrome redux.* Nephrology Dialysis Transplantation, 2014. **30**(1): p. 12-17.
48. Agrawal, S., et al., *Dyslipidaemia in nephrotic syndrome: mechanisms and treatment.* Nature reviews. Nephrology, 2018. **14**(1): p. 57-70.
49. Lu, W., et al., *Clinicopathological features and prognosis in patients with idiopathic membranous nephropathy with hypertension.* Exp Ther Med, 2020. **19**(4): p. 2615-2621.
50. Topham, P.S., et al., *Glomerular disease as a cause of isolated microscopic haematuria.* Q J Med, 1994. **87**(6): p. 329-35.
51. Li, S.J., et al., *Thromboembolic complications in membranous nephropathy patients with nephrotic syndrome-a prospective study.* Thromb Res, 2012. **130**(3): p. 501-5.
52. Lionaki, S., et al., *Venous thromboembolism in patients with membranous nephropathy.* Clinical journal of the American Society of Nephrology : CJASN, 2012. **7**(1): p. 43-51.
53. Fogo, A.B., et al., *AJKD Atlas of Renal Pathology: Membranous Nephropathy.* Am J Kidney Dis, 2015. **66**(3): p. e15-7.
54. Troyanov, S., et al., *Renal pathology in idiopathic membranous nephropathy: A new perspective.* Kidney International, 2006. **69**(9): p. 1641-1648.
55. Heymann, W., et al., *Production of nephrotic syndrome in rats by Freund's adjuvants and rat kidney suspensions.* Proc Soc Exp Biol Med, 1959. **100**(4): p. 660-4.
56. Couser, W.G., et al., *Experimental glomerulonephritis in the isolated perfused rat kidney.* The Journal of clinical investigation, 1978. **62**(6): p. 1275-1287.
57. Beck, L.H., et al., *M-Type Phospholipase A2 Receptor as Target Antigen in Idiopathic Membranous Nephropathy.* New England Journal of Medicine, 2009. **361**(1): p. 11-21.
58. Ancian, P., et al., *The human 180-kDa receptor for secretory phospholipases A2. Molecular cloning, identification of a secreted soluble form, expression, and chromosomal localization.* J Biol Chem, 1995. **270**(15): p. 8963-70.
59. Kinoshita, E., et al., *Activation of MAP kinase cascade induced by human pancreatic phospholipase A2in a human pancreatic cancer cell line.* FEBS Letters, 1997. **407**(3): p. 343-346.
60. Kanemasa, T., K. Hanasaki, and H. Arita, *Migration of vascular smooth muscle cells by phospholipase A2 via specific binding sites.* Biochimica et

Biophysica Acta (BBA) - Lipids and Lipid Metabolism, 1992. **1125**(2): p. 210-214.
61. Nomura, K., H. Fujita, and H. Arita, *Gene expression of pancreatic-type phospholipase-A2 in rat ovaries: stimulatory action on progesterone release.* Endocrinology, 1994. **135**(2): p. 603-9.
62. Hanasaki, K. and H. Arita, *Phospholipase A2 receptor: a regulator of biological functions of secretory phospholipase A2.* Prostaglandins Other Lipid Mediat, 2002. **68-69**: p. 71-82.
63. Hanasaki, K., et al., *Resistance to Endotoxic Shock in Phospholipase A₂ Receptor-deficient Mice *.* Journal of Biological Chemistry, 1997. **272**(52): p. 32792-32797.
64. Yokota, Y., et al., *Suppression of murine endotoxic shock by sPLA2 inhibitor, indoxam, through group IIA sPLA2-independent mechanisms.* Biochimica et biophysica acta, 1999. **1438**(2): p. 213-222.
65. Hanasaki, K. and H. Arita, *Characterization of a high affinity binding site for pancreatic-type phospholipase A2 in the rat. Its cellular and tissue distribution.* J Biol Chem, 1992. **267**(9): p. 6414-20.
66. East, L. and C.M. Isacke, *The mannose receptor family.* Biochim Biophys Acta, 2002. **1572**(2-3): p. 364-86.
67. Kao, L., et al., *Identification of the immunodominant epitope region in phospholipase A2 receptor-mediating autoantibody binding in idiopathic membranous nephropathy.* J Am Soc Nephrol, 2015. **26**(2): p. 291-301.
68. Fresquet, M., et al., *Identification of a major epitope recognized by PLA2R autoantibodies in primary membranous nephropathy.* J Am Soc Nephrol, 2015. **26**(2): p. 302-13.
69. Seitz-Polski, B., et al., *Epitope Spreading of Autoantibody Response to PLA2R Associates with Poor Prognosis in Membranous Nephropathy.* J Am Soc Nephrol, 2016. **27**(5): p. 1517-33.
70. University of California, S.C. *Human chr2:160766793-160949593 - UCSC Genome Browser v349*. 2009 25/05/2021]; Available from: https://genome.ucsc.edu/cgi-bin/hgTracks?db=hg19&lastVirtModeType=default&lastVirtModeExtraState=&virtModeType=default&virtMode=0&nonVirtPosition=&position=chr2%3A160766793%2D160949593&hgsid=595174069_KSHxo7VUtmphiOcZYF51guMCB6nD.
71. Fagerberg, L., et al., *Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics.* Mol Cell Proteomics, 2014. **13**(2): p. 397-406.
72. Coenen, M.J., et al., *Phospholipase A2 receptor (PLA2R1) sequence variants in idiopathic membranous nephropathy.* J Am Soc Nephrol, 2013. **24**(4): p. 677-83.
73. Stanescu, H.C., et al., *Risk HLA-DQA1 and PLA(2)R1 alleles in idiopathic membranous nephropathy.* N Engl J Med, 2011. **364**(7): p. 616-26.
74. Du, Y., et al., *The diagnosis accuracy of PLA2R-AB in the diagnosis of idiopathic membranous nephropathy: a meta-analysis.* PloS one, 2014. **9**(8): p. e104936-e104936.
75. Svobodova, B., et al., *Kidney biopsy is a sensitive tool for retrospective diagnosis of PLA2R-related membranous nephropathy.* Nephrol Dial Transplant, 2013. **28**(7): p. 1839-44.

76. Bech, A.P., et al., *Association of anti-PLA(2)R antibodies with outcomes after immunosuppressive therapy in idiopathic membranous nephropathy.* Clin J Am Soc Nephrol, 2014. **9**(8): p. 1386-92.
77. Hofstra, J.M., et al., *Anti-phospholipase A(2) receptor antibodies correlate with clinical status in idiopathic membranous nephropathy.* Clin J Am Soc Nephrol, 2011. **6**(6): p. 1286-91.
78. Hoxha, E., et al., *Phospholipase A2 receptor autoantibodies and clinical outcome in patients with primary membranous nephropathy.* J Am Soc Nephrol, 2014. **25**(6): p. 1357-66.
79. Ruggenenti, P., et al., *Anti-Phospholipase A2 Receptor Antibody Titer Predicts Post-Rituximab Outcome of Membranous Nephropathy.* J Am Soc Nephrol, 2015. **26**(10): p. 2545-58.
80. Jullien, P., et al., *Anti-phospholipase A2 receptor antibody levels at diagnosis predicts spontaneous remission of idiopathic membranous nephropathy.* Clinical Kidney Journal, 2017. **10**(2): p. 209-214.
81. Beck, L.H., Jr., et al., *Rituximab-induced depletion of anti-PLA2R autoantibodies predicts response in membranous nephropathy.* J Am Soc Nephrol, 2011. **22**(8): p. 1543-50.
82. van de Logt, A.E., et al., *Immunological remission in PLA2R-antibody-associated membranous nephropathy: cyclophosphamide versus rituximab.* Kidney Int, 2018. **93**(4): p. 1016-1017.
83. van de Logt, A.E., et al., *The anti-PLA2R antibody in membranous nephropathy: what we know and what remains a decade after its discovery.* Kidney Int, 2019. **96**(6): p. 1292-1302.
84. Hoxha, E., et al., *An Indirect Immunofluorescence Method Facilitates Detection of Thrombospondin Type 1 Domain–Containing 7A–Specific Antibodies in Membranous Nephropathy.* Journal of the American Society of Nephrology, 2017. **28**(2): p. 520-531.
85. Debiec, H. and P. Ronco, *PLA2R autoantibodies and PLA2R glomerular deposits in membranous nephropathy.* N Engl J Med, 2011. **364**(7): p. 689-90.
86. van de Logt, A.E., J.M. Hofstra, and J.F. Wetzels, *Serum anti-PLA2R antibodies can be initially absent in idiopathic membranous nephropathy: seroconversion after prolonged follow-up.* Kidney Int, 2015. **87**(6): p. 1263-4.
87. Ramachandran, R., et al., *Serial monitoring of anti-PLA₂R in initial PLA₂R-negative patients with primary membranous nephropathy.* Kidney International, 2015. **88**(5): p. 1198-1199.
88. De Vriese, A.S., et al., *A Proposal for a Serology-Based Approach to Membranous Nephropathy.* J Am Soc Nephrol, 2017. **28**(2): p. 421-430.
89. Borza, D.B., et al., *Mouse models of membranous nephropathy: the road less travelled by.* Am J Clin Exp Immunol, 2013. **2**(2): p. 135-45.
90. Seitz-Polski, B., et al., *Cross-reactivity of anti-PLA2R1 autoantibodies to rabbit and mouse PLA2R1 antigens and development of two novel ELISAs with different diagnostic performances in idiopathic membranous nephropathy.* Biochimie, 2015. **118**: p. 104-15.
91. Lv, J., et al., *Interaction between PLA2R1 and HLA-DQA1 variants associates with anti-PLA2R antibodies and membranous nephropathy.* J Am Soc Nephrol, 2013. **24**(8): p. 1323-9.
92. Saeed, M., et al., *PLA2R-associated membranous glomerulopathy is modulated by common variants in PLA2R1 and HLA-DQA1 genes.* Genes Immun, 2014. **15**(8): p. 556-61.

93. Kanigicherla, D., et al., *Anti-PLA2R antibodies measured by ELISA predict long-term outcome in a prevalent population of patients with idiopathic membranous nephropathy.* Kidney Int, 2013. **83**(5): p. 940-8.
94. Ramachandran, R., et al., *PLA2R antibodies, glomerular PLA2R deposits and variations in PLA2R1 and HLA-DQA1 genes in primary membranous nephropathy in South Asians.* Nephrol Dial Transplant, 2016. **31**(9): p. 1486-93.
95. Cui, Z., et al., *MHC Class II Risk Alleles and Amino Acid Residues in Idiopathic Membranous Nephropathy.* J Am Soc Nephrol, 2017. **28**(5): p. 1651-1664.
96. Stehle, T., et al., *Phospholipase A2 receptor and sarcoidosis-associated membranous nephropathy.* Nephrol Dial Transplant, 2015. **30**(6): p. 1047-50.
97. Li, Y., et al., *Single-nucleotide polymorphisms in the PLA2R1 gene are associated with systemic lupus erythematosus and lupus nephritis in a Chinese Han population.* Immunol Res, 2016. **64**(1): p. 324-8.
98. Tomas, N.M., et al., *Thrombospondin type-1 domain-containing 7A in idiopathic membranous nephropathy.* N Engl J Med, 2014. **371**(24): p. 2277-2287.
99. Gödel, M., F. Grahammer, and T.B. Huber, *Thrombospondin type-1 domain-containing 7A in idiopathic membranous nephropathy.* N Engl J Med, 2015. **372**(11): p. 1073.
100. Seifert, L., et al., *The Most N-Terminal Region of THSD7A Is the Predominant Target for Autoimmunity in THSD7A-Associated Membranous Nephropathy.* Journal of the American Society of Nephrology, 2018. **29**(5): p. 1536-1548.
101. Ren, S., et al., *An update on clinical significance of use of THSD7A in diagnosing idiopathic membranous nephropathy: a systematic review and meta-analysis of THSD7A in IMN.* Renal Failure, 2018. **40**(1): p. 306-313.
102. Larsen, C.P., L.N. Cossey, and L.H. Beck, *THSD7A staining of membranous glomerulopathy in clinical practice reveals cases with dual autoantibody positivity.* Mod Pathol, 2016. **29**(4): p. 421-6.
103. Fresquet, M., et al., *Autoantigens PLA2R and THSD7A in membranous nephropathy share a common epitope motif in the N-terminal domain.* Journal of Autoimmunity, 2020. **106**: p. 102308.
104. Sethi, S., et al., *Exostosin 1/Exostosin 2-Associated Membranous Nephropathy.* J Am Soc Nephrol, 2019. **30**(6): p. 1123-1136.
105. Ravindran, A., et al., *In Patients with Membranous Lupus Nephritis, Exostosin-Positivity and Exostosin-Negativity Represent Two Different Phenotypes.* Journal of the American Society of Nephrology, 2021. **32**(3): p. 695-706.
106. Sethi, S., et al., *Neural epidermal growth factor-like 1 protein (NELL-1) associated membranous nephropathy.* Kidney Int, 2020. **97**(1): p. 163-174.
107. Caza, T.N., et al., *NELL1 is a target antigen in malignancy-associated membranous nephropathy.* Kidney International, 2021. **99**(4): p. 967-976.
108. Sethi, S., et al., *Semaphorin 3B–associated membranous nephropathy is a distinct type of disease predominantly present in pediatric patients.* Kidney International, 2020. **98**(5): p. 1253-1264.
109. Al-Rabadi, L.F., et al., *Serine Protease HTRA1 as a Novel Target Antigen in Primary Membranous Nephropathy.* J Am Soc Nephrol, 2021.
110. Sethi, S., et al., *Protocadherin 7–Associated Membranous Nephropathy.* Journal of the American Society of Nephrology, 2021. **32**(5): p. 1249-1261.

111. Le Quintrec, M., et al., *Contactin-1 is a novel target antigen in membranous nephropathy associated with chronic inflammatory demyelinating polyneuropathy.* Kidney Int, 2021.
112. Witte, A.S. and J.F. Burke, *Membranous glomerulonephritis associated with chronic progressive demyelinating neuropathy.* Neurology, 1987. **37**(2): p. 342-342.
113. Hashimoto, Y., et al., *Chronic Inflammatory Demyelinating Polyneuropathy With Concurrent Membranous Nephropathy: An Anti-paranode and Podocyte Protein Antibody Study and Literature Survey.* Frontiers in neurology, 2018. **9**: p. 997-997.
114. Delmont, E., et al., *Antibodies against the node of Ranvier: a real-life evaluation of incidence, clinical features and response to treatment based on a prospective analysis of 1500 sera.* Journal of Neurology, 2020. **267**(12): p. 3664-3672.
115. Cortese, A., et al., *Antibodies to neurofascin, contactin-1, and contactin-associated protein 1 in CIDP.* Clinical relevance of IgG isotype, 2020. **7**(1): p. e639.
116. Caza, T.N., et al., *Neural cell adhesion molecule 1 is a novel autoantigen in membranous lupus nephritis.* Kidney International, 2020.
117. Bruschi, M., et al., *Direct characterization of target podocyte antigens and auto-antibodies in human membranous glomerulonephritis: Alfa-enolase and borderline antigens.* Journal of Proteomics, 2011. **74**(10): p. 2008-2017.
118. Kimura, Y., et al., *Circulating antibodies to α-enolase and phospholipase A2 receptor and composition of glomerular deposits in Japanese patients with primary or secondary membranous nephropathy.* Clinical and Experimental Nephrology, 2017. **21**(1): p. 117-126.
119. Prunotto, M., et al., *Autoimmunity in Membranous Nephropathy Targets Aldose Reductase and SOD2.* Journal of the American Society of Nephrology, 2010. **21**(3): p. 507-519.
120. Murtas, C. and G.M. Ghiggeri, *Membranous glomerulonephritis: histological and serological features to differentiate cancer-related and non-related forms.* Journal of Nephrology, 2016. **29**(4): p. 469-478.
121. Glassock, R.J., *Pathogenesis of Membranous Nephropathy: A new paradigm in evolution*, in *New Insights into Glomerulonephritis. Pathogenesis and treatment.*, N. Chen, Editor. 2013, Karger.
122. Nangaku, M., S.J. Shankland, and W.G. Couser, *Cellular Response to Injury in Membranous Nephropathy.* Journal of the American Society of Nephrology, 2005. **16**(5): p. 1195-1204.
123. Liu, W., et al., *Immunological Pathogenesis of Membranous Nephropathy: Focus on PLA2R1 and Its Role.* Frontiers in Immunology, 2019. **10**(1809).
124. Salant, D.J., *In search of the elusive membranous nephropathy antigen.* Nephron Physiol, 2009. **112**(1): p. p11-2.
125. Glassock, R.J., *Human idiopathic membranous nephropathy--a mystery solved?* N Engl J Med, 2009. **361**(1): p. 81-3.
126. Vasmant, D., et al., *Familial idiopathic membranous glomerulonephritis.* Int J Pediatr Nephrol, 1984. **5**(4): p. 193-6.
127. Bockenhauer, D., et al., *Familial membranous nephropathy: an X-linked genetic susceptibility?* Nephron Clin Pract, 2008. **108**(1): p. c10-5.

128. Dyer, P.A., et al., *HLA antigen and gene polymorphisms and haplotypes established by family studies in membranous nephropathy.* Nephrol Dial Transplant, 1992. **7 Suppl 1**: p. 42-7.
129. Elshihabi, I., C.I. Kaye, and A. Brzowski, *Membranous nephropathy in two human leukocyte antigen-identical brothers.* J Pediatr, 1993. **123**(6): p. 940-2.
130. Grcevska, L. and M. Polenakovic, *Idiopathic membranous nephropathy (IMN) in two HLA-identical brothers with different outcome of the disease.* Clin Nephrol, 1999. **52**(3): p. 194-6.
131. Izzi, C., et al., *Familial aggregation of primary glomerulonephritis in an Italian population isolate: Valtrompia study.* Kidney Int, 2006. **69**(6): p. 1033-40.
132. Maccario, M., et al., *Idiopathic membranous nephropathy in two siblings.* Nephrol Dial Transplant, 1995. **10**(1): p. 108-10.
133. Meroni, M., et al., *Two brothers with idiopathic membranous nephropathy and familial sensorineural deafness.* Am J Kidney Dis, 1990. **15**(3): p. 269-72.
134. Muller, C., et al., *Familial membranous glomerulopathy, toxic exposure and/or genetic sensibility?* Clin Nephrol, 2008. **70**(5): p. 422-3.
135. Sato, K., et al., *Idiopathic membranous nephropathy in two brothers.* Nephron, 1987. **46**(2): p. 174-8.
136. Scolari, F., et al., *Familial membranous nephropathy.* J Nephrol, 1998. **11**(1): p. 35-9.
137. Short, C.D., et al., *Familial membranous nephropathy.* Br Med J (Clin Res Ed), 1984. **289**(6457): p. 1500.
138. Vangelista, A., R. Tazzari, and V. Bonomini, *Idiopathic membranous nephropathy in 2 twin brothers.* Nephron, 1988. **50**(1): p. 79-80.
139. Tao, T., et al., *Identical twins with idiopathic membranous nephropathy.* Journal of Nephrology, 2021. **34**(2): p. 597-601.
140. Downie, M.L., et al., *Identification of a Locus on the X Chromosome Linked to Familial Membranous Nephropathy.* Kidney International Reports, 2021.
141. Pierides, A.M., et al., *Idiopathic membranous nephropathy.* Q J Med, 1977. **46**(182): p. 163-77.
142. Mezzano, S., et al., *Idiopathic membranous nephropathy, associated with HLA-DRw3 and not related to monocyte-phagocyte system Fc receptor dysfunction, in father and son.* Nephron, 1991. **58**(3): p. 320-4.
143. Kitsios, G.D. and E. Zintzaras, *Genome-wide association studies: hypothesis-"free" or "engaged"?* Transl Res, 2009. **154**(4): p. 161-4.
144. Karalliedde, J. and L. Gnudi, *Diabetes mellitus, a complex and heterogeneous disease, and the role of insulin resistance as a determinant of diabetic kidney disease.* Nephrol Dial Transplant, 2016. **31**(2): p. 206-13.
145. Segelmark, M., *Genes that link nephritis to autoantibodies and innate immunity.* N Engl J Med, 2011. **364**(7): p. 679-80.
146. Bomback, A.S. and A.G. Gharavi, *Can Genetics Risk-Stratify Patients with Membranous Nephropathy?* Journal of the American Society of Nephrology, 2013. **24**(8): p. 1190-1192.
147. Sekula, P., et al., *Genetic risk variants for membranous nephropathy: extension of and association with other chronic kidney disease aetiologies.* Nephrol Dial Transplant, 2017. **32**(2): p. 325-332.
148. Fernando, M.M. and T.J. Vyse, *Risk alleles in idiopathic membranous nephropathy.* N Engl J Med, 2011. **364**(21): p. 2072; author reply 2073-4.
149. Kiryluk, K., *Risk alleles in idiopathic membranous nephropathy.* N Engl J Med, 2011. **364**(21): p. 2072-3; author reply 2073-4.

150. Marchini, J. and B. Howie, *Genotype imputation for genome-wide association studies.* Nat Rev Genet, 2010. **11**(7): p. 499-511.
151. Spencer, C.C., et al., *Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip.* PLoS Genet, 2009. **5**(5): p. e1000477.
152. Xu, X., et al., *Long-term exposure to air pollution and increased risk of membranous nephropathy in China.* Journal of the American Society of Nephrology, 2016. **27**(12): p. 3739-3746.
153. Kwon, J.M. and A.M. Goate, *The candidate gene approach.* Alcohol Res Health, 2000. **24**(3): p. 164-8.
154. Tabor, H.K., N.J. Risch, and R.M. Myers, *Candidate-gene approaches for studying complex genetic traits: practical considerations.* Nat Rev Genet, 2002. **3**(5): p. 391-7.
155. Liu, Y.H., et al., *Association of phospholipase A2 receptor 1 polymorphisms with idiopathic membranous nephropathy in Chinese patients in Taiwan.* J Biomed Sci, 2010. **17**(1): p. 81.
156. Wang, F., et al., *PLA2R1 and HLA-DQA1 gene variations in idiopathic membranous nephropathy in South China.* Ann Acad Med Singap, 2021. **50**(1): p. 33-41.
157. Fan, S., et al., *The association between variants in PLA2R and HLA-DQA1 and renal outcomes in patients with primary membranous nephropathy in Western China.* BMC Medical Genomics, 2021. **14**(1): p. 123.
158. Le, W.B., et al., *HLA-DRB1*15:01 and HLA-DRB3*02:02 in PLA2R-Related Membranous Nephropathy.* J Am Soc Nephrol, 2017. **28**(5): p. 1642-1650.
159. Wang, W., et al., *Interaction between PLA2R1 and HLA-DQA1 variants contributes to the increased genetic susceptibility to membranous nephropathy in Western China.* Nephrology, 2019. **24**(9): p. 919-925.
160. Tian, C.X., et al., *Association of SNPs in PLA2R1 with idiopathic and secondary membranous nephropathy in two Chinese cohorts.* British Journal of Biomedical Science, 2020. **77**(1): p. 24-28.
161. Cui, G., et al., *Development of a high resolution melting method for genotyping of risk HLA-DQA1 and PLA2R1 alleles and ethnic distribution of these risk alleles.* Gene, 2013. **514**(2): p. 125-30.
162. Akiyama, S., et al., *Prevalence of anti-phospholipase A2 receptor antibodies in Japanese patients with membranous nephropathy.* Clin Exp Nephrol, 2015. **19**(4): p. 653-60.
163. Hinrichs, A.S., et al., *The UCSC Genome Browser Database: update 2006.* Nucleic Acids Res, 2006. **34**(Database issue): p. D590-8.
164. Kaga, H., et al., *Analysis of PLA2R1 and HLA-DQA1 sequence variants in Japanese patients with idiopathic and secondary membranous nephropathy.* Clin Exp Nephrol, 2018. **22**(2): p. 275-282.
165. Thiri, M., et al., *High-density Association Mapping and Interaction Analysis of PLA2R1 and HLA Regions with Idiopathic Membranous Nephropathy in Japanese.* Sci Rep, 2016. **6**: p. 38189.
166. Latt, K.Z., et al., *Identification of a two-SNP PLA2R1 Haplotype and HLA-DRB1 Alleles as Primary Risk Associations in Idiopathic Membranous Nephropathy.* Scientific Reports, 2018. **8**(1): p. 15576.
167. OECD, *Chapter 3: Health check ups in Japan*, in *OECD Reviews of Public Health: Japan. A healthier tomorrow.* 2019, OECD Publishing: Paris. p. 208.

168. Imai, E., et al., *Kidney Disease Screening Program in Japan: History, Outcome, and Perspectives.* Clinical Journal of the American Society of Nephrology, 2007. **2**(6): p. 1360-1366.
169. Le, W.-B., et al., *HLA Alleles and Prognosis of PLA2R-Related Membranous Nephropathy.* Clinical Journal of the American Society of Nephrology, 2021: p. CJN.18021120.
170. Lowe, M., et al., *Associations between human leukocyte antigens and renal function.* Scientific reports, 2021. **11**(1): p. 3158-3158.
171. Xie, J., et al., *The genetic architecture of membranous nephropathy and its potential to improve non-invasive diagnosis.* Nature Communications, 2020. **11**(1): p. 1600.
172. Berchtold, L., et al., *HLA-D and PLA2R1 risk alleles associate with recurrent primary membranous nephropathy in kidney transplant recipients.* Kidney International, 2021. **99**(3): p. 671-685.
173. Ponticelli, C. and R.J. Glassock, *Glomerular Diseases: Membranous Nephropathy—A Modern View.* Clinical Journal of the American Society of Nephrology, 2014. **9**(3): p. 609-616.
174. Salant, D.J., *Genetic variants in membranous nephropathy: perhaps a perfect storm rather than a straightforward conformeropathy?* J Am Soc Nephrol, 2013. **24**(4): p. 525-8.
175. Horton, R., et al., *Variation analysis and gene annotation of eight MHC haplotypes: The MHC Haplotype Project.* Immunogenetics, 2008. **60**(1): p. 1-18.
176. Wikipedia. *HLA A1-B8-DR3-DQ2*. 2021 [cited 2022 18/05/2022]; Available from: https://en.wikipedia.org/wiki/HLA_A1-B8-DR3-DQ2.
177. Manabe, K., et al., *Human leukocyte antigen A1-B8-DR3-DQ2-DPB1*0401 extended haplotype in autoimmune hepatitis.* Hepatology, 1993. **18**(6): p. 1334-1337.
178. Winchester, A.M. *Genetics*. 2020 08/06/2021]; Available from: https://www.britannica.com/science/genetics.
179. European Molecular Biology Laboratory - European Bioinformatics Institute. *What is genetic variation*. 2021; Available from: https://www.ebi.ac.uk/training/online/courses/human-genetic-variation-introduction/what-is-genetic-variation/.
180. National Institutes of Health, U. *Understanding Human Genetic Variation*. Biological Sciences Curriculum Study. NIH Curriculum Supplement Series 2007 [cited 2021 08/06/2021]; Available from: https://www.ncbi.nlm.nih.gov/books/NBK20363/.
181. Eichler, E.E., *Genetic Variation, Comparative Genomics, and the Diagnosis of Disease.* The New England journal of medicine, 2019. **381**(1): p. 64-74.
182. Alkan, C., B.P. Coe, and E.E. Eichler, *Genome structural variation discovery and genotyping.* Nature Reviews Genetics, 2011. **12**(5): p. 363-376.
183. Knibbe, C., et al., *A Long-Term Evolutionary Pressure on the Amount of Noncoding DNA.* Molecular Biology and Evolution, 2007. **24**(10): p. 2344-2353.
184. Gullapalli, R.R., et al., *Next generation sequencing in clinical medicine: Challenges and lessons for pathology and biomedical informatics.* J Pathol Inform, 2012. **3**: p. 40.

185. Hehir-Kwa, J.Y., et al., *Towards a European consensus for reporting incidental findings during clinical NGS testing.* Eur J Hum Genet, 2015. **23**(12): p. 1601-6.
186. Schochetman, G., C.Y. Ou, and W.K. Jones, *Polymerase chain reaction.* J Infect Dis, 1988. **158**(6): p. 1154-7.
187. Wilton, S., *Long-range PCR*, in *eLS*.
188. Scientific, T.F. *Long-Range PCR Enzymes & Master Mixes—Thermo Scientific*. 2017 [cited 2017 09/06/2017]; Available from: https://www.thermofisher.com/uk/en/home/brands/thermo-scientific/molecular-biology/thermo-scientific-pcr/thermo-scientific-pcr-enzymes-master-mixes/long-range-pcr-master-mixes-thermo-scientific.html.
189. QIAGEN. *QIAGEN LongRange PCR Kit - QIAGEN Online Shop*. 2017 [cited 2017 09/06/2017]; Available from: https://www.qiagen.com/us/shop/pcr/end-point-pcr-enzymes-and-kits/regular-pcr/qiagen-longrange-pcr-kit/#orderinginformation.
190. Jia, H., et al., *Long-range PCR in next-generation sequencing: comparison of six enzymes and evaluation on the MiSeq sequencer.* Scientific Reports, 2014. **4**(1): p. 5737.
191. Mullis, K.B. and F.A. Faloona, *Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction.* Methods Enzymol, 1987. **155**: p. 335-50.
192. Sanger, F., S. Nicklen, and A.R. Coulson, *DNA sequencing with chain-terminating inhibitors.* Proc Natl Acad Sci U S A, 1977. **74**(12): p. 5463-7.
193. Sham, P., et al., *DNA Pooling: a tool for large-scale association studies.* Nat Rev Genet, 2002. **3**(11): p. 862-71.
194. Amos, C.I., M.L. Frazier, and W. Wang, *DNA pooling in mutation detection with reference to sequence analysis.* American journal of human genetics, 2000. **66**(5): p. 1689-1692.
195. Sommer, S.S., et al., *A Novel Method for Detecting Point Mutations or Polymorphisms and Its Application to Population Screening for Carriers of Phenylketonuria.* Mayo Clinic Proceedings, 1989. **64**(11): p. 1361-1372.
196. He, C., J. Holme, and J. Anthony, *SNP genotyping: the KASP assay.* Methods Mol Biol, 2014. **1145**: p. 75-86.
197. Samorodnitsky, E., et al., *Evaluation of Hybridization Capture Versus Amplicon-Based Methods for Whole-Exome Sequencing.* Hum Mutat, 2015. **36**(9): p. 903-14.
198. Mao, X., B.D. Young, and Y.-J. Lu, *The application of single nucleotide polymorphism microarrays in cancer research.* Current genomics, 2007. **8**(4): p. 219-228.
199. Waddell, N., *Microarray-based DNA profiling to study genomic aberrations.* IUBMB Life, 2008. **60**(7): p. 437-440.
200. Illumina. *Microarray Data Analysis Workflows*. 2016 [cited 2021 01/11/2021]; Available from: https://emea.illumina.com/content/dam/illumina-marketing/documents/products/technotes/technote_array_analysis_workflows.pdf.
201. Smith, M.L., et al., *illuminaio: An open source IDAT parsing tool for Illumina microarrays.* F1000Research, 2013. **2**: p. 264-264.
202. Ansorge, W.J., *Next-generation DNA sequencing techniques.* N Biotechnol, 2009. **25**(4): p. 195-203.

203. Reuter, J.A., D.V. Spacek, and M.P. Snyder, *High-throughput sequencing technologies.* Molecular cell, 2015. **58**(4): p. 586-597.
204. Shendure, J. and H. Ji, *Next-generation DNA sequencing.* Nat Biotechnol, 2008. **26**(10): p. 1135-45.
205. Mardis, E.R., *Next-generation DNA sequencing methods.* Annu Rev Genomics Hum Genet, 2008. **9**: p. 387-402.
206. Flint, J., *GWAS.* Current Biology, 2013. **23**(7): p. R265-6.
207. Viennas, E., et al., *Population-ethnic group specific genome variation allele frequency data: A querying and visualization journey.* Genomics, 2012. **100**(2): p. 93-101.
208. Bush, W.S. and J.H. Moore, *Chapter 11: Genome-wide association studies.* PLoS computational biology, 2012. **8**(12): p. e1002822-e1002822.
209. Slatkin, M., *Linkage disequilibrium — understanding the evolutionary past and mapping the medical future.* Nature Reviews Genetics, 2008. **9**(6): p. 477-485.
210. Wolf, J., E.D. Brodie, and M.J. Wade, *Epistasis and the evolutionary process*. 2000, USA: Oxford University Press.
211. Cordell, H.J., *Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans.* Human Molecular Genetics, 2002. **11**(20): p. 2463-2468.
212. Ritchie, M.D., *Finding the epistasis needles in the genome-wide haystack.* Epistasis, 2015: p. 19-33.
213. Li, Y., et al., *Genotype imputation.* Annu Rev Genomics Hum Genet, 2009. **10**: p. 387-406.
214. Burdick, J.T., et al., *In silico method for inferring genotypes in pedigrees.* Nature genetics, 2006. **38**(9): p. 1002-1004.
215. Browning, B.L., Y. Zhou, and S.R. Browning, *A One-Penny Imputed Genome from Next-Generation Reference Panels.* Am J Hum Genet, 2018. **103**(3): p. 338-348.
216. Browning, B.L. and S.R. Browning, *A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals.* American journal of human genetics, 2009. **84**(2): p. 210-223.
217. Horton, R., et al., *Gene map of the extended human MHC.* Nat Rev Genet, 2004. **5**(12): p. 889-99.
218. Liu, P., et al., *Benchmarking the Human Leukocyte Antigen Typing Performance of Three Assays and Seven Next-Generation Sequencing-Based Algorithms.* Frontiers in Immunology, 2021. **12**(840).
219. Choo, S.Y., *The HLA system: genetics, immunology, clinical testing, and clinical implications.* Yonsei medical journal, 2007. **48**(1): p. 11-23.
220. Hirata, J., et al., *Genetic and phenotypic landscape of the major histocompatibilty complex region in the Japanese population.* Nat Genet, 2019. **51**(3): p. 470-480.
221. Erlich, H., *HLA DNA typing: past, present, and future.* Tissue Antigens, 2012. **80**(1): p. 1-11.
222. Dendrou, C.A., et al., *HLA variation and disease.* Nat Rev Immunol, 2018. **18**(5): p. 325-339.
223. Naito, T., et al., *A deep learning method for HLA imputation and trans-ethnic MHC fine-mapping of type 1 diabetes.* Nature Communications, 2021. **12**(1): p. 1639.

224. Ritari, J., et al., *Increasing accuracy of HLA imputation by a population-specific reference panel in a FinnGen biobank cohort.* NAR Genomics and Bioinformatics, 2020. **2**(2).
225. Jia, X., et al., *Imputing amino acid polymorphisms in human leukocyte antigens.* PLoS One, 2013. **8**(6): p. e64683.
226. Karnes, J.H., et al., *Comparison of HLA allelic imputation programs.* PLOS ONE, 2017. **12**(2): p. e0172444.
227. Lewis, C.M. and E. Vassos, *Polygenic risk scores: from research tools to clinical instruments.* Genome Medicine, 2020. **12**(1): p. 44.
228. Wray, N.R., et al., *Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression.* Nat Genet, 2018. **50**(5): p. 668-681.
229. Mavaddat, N., et al., *Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes.* Am J Hum Genet, 2019. **104**(1): p. 21-34.
230. Khera, A.V., et al., *Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations.* Nature genetics, 2018. **50**(9): p. 1219-1224.
231. Musliner, K.L., et al., *Association of Polygenic Liabilities for Major Depression, Bipolar Disorder, and Schizophrenia With Risk for Depression in the Danish Population.* JAMA Psychiatry, 2019. **76**(5): p. 516-525.
232. Lewis, C.M. and S.P. Hagenaars, *Progressing Polygenic Medicine in Psychiatry Through Electronic Health Records.* JAMA Psychiatry, 2019. **76**(5): p. 470-472.
233. Torkamani, A., N.E. Wineinger, and E.J. Topol, *The personal and clinical utility of polygenic risk scores.* Nat Rev Genet, 2018. **19**(9): p. 581-590.
234. Polderman, T.J., et al., *Meta-analysis of the heritability of human traits based on fifty years of twin studies.* Nat Genet, 2015. **47**(7): p. 702-9.
235. Tippin-Davis, B. *Clinical Implementation of a Polygenic Risk Score (PRS) for Breast Cancer*. Clinical Trials 2018 [cited 2021 11/06/2021]; Available from: https://clinicaltrials.gov/ct2/show/NCT03688204.
236. Gonzalez-Perez, A., et al., *Computational approaches to identify functional genetic variants in cancer genomes.* Nature methods, 2013. **10**(8): p. 723-729.
237. Kent, W.J., et al., *The human genome browser at UCSC.* Genome research, 2002. **12**(6): p. 996-1006.
238. McLaren, W., et al., *The Ensembl Variant Effect Predictor.* Genome biology, 2016. **17**(1): p. 122-122.
239. Woolfe, A., et al., *Highly conserved non-coding sequences are associated with vertebrate development.* PLoS biology, 2005. **3**(1): p. e7-e7.
240. Siavrienė, E. and V. Kucinskas, *The most common technologies and tools for functional genome analysis.* Acta medica Lituanica, 2017. **24**: p. 1-11.
241. Thermo Fisher Scientific. *Methods for Detecting Protein–DNA Interactions*. 2010 [cited 2021 26/10/2021]; Available from: https://www.thermofisher.com/uk/en/home/life-science/protein-biology/protein-biology-learning-center/protein-biology-resource-library/pierce-protein-methods/methods-detecting-protein-dna-interactions.html.
242. Fried, M.G., *Measurement of protein-DNA interaction parameters by electrophoresis mobility shift assay.* Electrophoresis, 1989. **10**(5-6): p. 366-76.
243. Hellman, L.M. and M.G. Fried, *Electrophoretic mobility shift assay (EMSA) for detecting protein-nucleic acid interactions.* Nat Protoc, 2007. **2**(8): p. 1849-61.

244. Alves, C. and C. Cunha, *Electrophoretic Mobility Shift Assay: Analyzing Protein – Nucleic Acid Interactions*, in *Gel Electrophoresis - Advanced Techniques*, S. Magdeldin, Editor. 2012, InTech.
245. !!! INVALID CITATION !!! [222].
246. Auton, A., et al., *A global reference for human genetic variation.* Nature, 2015. **526**(7571): p. 68-74.
247. International Genome Sample Resource. *Populations in the International Genome Sample resource*. 2021 [cited 2021 17/11/2021]; Available from: https://www.internationalgenome.org/data-portal/population.
248. Team, R.C., *R: A language and environment for statistical computing*. 2013: R Foundation for Statistical Computing, Vienna, Austria.
249. Stefanescu, C., V.W. Berger, and S.L. Hershberger, *Yates' Correction*, in *Encyclopedia of Statistics in Behavioral Science*. 2005.
250. Pruim, R.J., et al., *LocusZoom: regional visualization of genome-wide association scan results.* Bioinformatics (Oxford, England), 2010. **26**(18): p. 2336-2337.
251. Excoffier, L. and M. Slatkin, *Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population.* Mol Biol Evol, 1995. **12**(5): p. 921-7.
252. Chang, C.C., et al., *Second-generation PLINK: rising to the challenge of larger and richer datasets.* GigaScience, 2015. **4**(1).
253. Hinrichs, A.S., et al., *UCSC Data Integrator and Variant Annotation Integrator.* Bioinformatics (Oxford, England), 2016. **32**(9): p. 1430-1432.
254. !!! INVALID CITATION !!! [229].
255. O'Leary, N.A., et al., *Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation.* Nucleic Acids Res, 2016. **44**(D1): p. D733-45.
256. Sherry, S.T., et al., *dbSNP: the NCBI database of genetic variation.* Nucleic acids research, 2001. **29**(1): p. 308-311.
257. Li, Q. and K. Wang, *InterVar: Clinical Interpretation of Genetic Variants by the 2015 ACMG-AMP Guidelines.* American journal of human genetics, 2017. **100**(2): p. 267-280.
258. Landrum, M.J., et al., *ClinVar: improving access to variant interpretations and supporting evidence.* Nucleic Acids Res, 2018. **46**(D1): p. D1062-d1067.
259. Amberger, J.S., et al., *OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders.* Nucleic Acids Res, 2015. **43**(Database issue): p. D789-98.
260. Karczewski, K.J., et al., *The ExAC browser: displaying reference data information from over 60 000 exomes.* Nucleic acids research, 2017. **45**(D1): p. D840-D845.
261. Desmet, F.-O., et al., *Human Splicing Finder: an online bioinformatics tool to predict splicing signals.* Nucleic acids research, 2009. **37**(9): p. e67-e67.
262. Davis, C.A., et al., *The Encyclopedia of DNA elements (ENCODE): data portal update.* Nucleic Acids Res, 2018. **46**(D1): p. D794-d801.
263. Matys, V., et al., *TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes.* Nucleic Acids Res, 2006. **34**(Database issue): p. D108-10.
264. Wingender, E., et al., *The TRANSFAC system on gene expression regulation.* Nucleic Acids Res, 2001. **29**(1): p. 281-3.

265. Hurst, R., et al., *The TNT® T7 Quick Coupled Transcription/Translation System.* reactions, 1996. **1**: p. 5.
266. Illumina. *Microarray kits for genotyping and epigenetic analysis*. 2018 [cited 2018 14/11/2018]; Available from: https://emea.illumina.com/products/by-type/microarray-kits.html.
267. Ye, J., et al., *Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction.* BMC bioinformatics, 2012. **13**(1): p. 1-11.
268. National Centre for biotechnology information. *Primer-BLAST. A tool for finding specific primers*. 2018 [cited 2018 19/11/2019]; Available from: https://www.ncbi.nlm.nih.gov/tools/primer-blast/.
269. Insightful Science, *SnapGene software*. 2017, Insightful Science.
270. Sigma-Aldrich. *High Pure PCR Product Purification Kit HPPCRPKRO Roche*. 2018 20/11/2018]; Available from: https://www.sigmaaldrich.com/GB/en/product/roche/hppcrpkro.
271. Eurofins Genomics. *Sample Submission Guides*. 2018 [cited 2018 20/11/2018]; Available from: https://eurofinsgenomics.eu/en/custom-dna-sequencing/additional-services/sample-submission.
272. Griekspoor, A. and T. Groothuis, *4Peaks: a program that helps molecular biologists to visualize and edit their DNA sequence files v1. 7*. 2005, Available at: h ttp://nucleobytes. com/index. php/4pea ks (accessed 12 ….
273. Chang, C., *PLINK v2.0*. 2018.
274. Purcell, S., et al., *PLINK: a tool set for whole-genome association and population-based linkage analyses.* Am J Hum Genet, 2007. **81**(3): p. 559-75.
275. Purcell, S. p. PLINK.
276. Chang, C.C., et al., *Second-generation PLINK: rising to the challenge of larger and richer datasets.* Gigascience, 2015. **4**: p. 7.
277. Chang, C. *PLINK File format reference*. 2021 [cited 2021 29/06/2021]; Available from: https://www.cog-genomics.org/plink/2.0/formats.
278. Bozeman, M., *NP & Variation Suite ™ (Version 8.8.1)* 2017, Golden Helix, Inc.
279. Li, H., *A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data.* Bioinformatics (Oxford, England), 2011. **27**(21): p. 2987-2993.
280. Danecek, P., et al., *Twelve years of SAMtools and BCFtools.* Gigascience, 2021. **10**(2).
281. Li, H., P. Danecek, and J. Marshall. *The Variant Call Format (VCF) Version 4.2 Specification*. 2021; Available from: https://samtools.github.io/hts-specs/VCFv4.2.pdf.
282. Browning, S.R. and B.L. Browning, *Haplotype phasing: existing methods and new developments.* Nature reviews. Genetics, 2011. **12**(10): p. 703-714.
283. Cheshire, C., *Bioinformatic Investigations Into the Genetic Architecture of Renal Disorders.* , in *Department of Medicine*. 2019, University College London: London.
284. Koboldt, D. *Tutorial: Allele coding conversion*. 2017 [cited 2021 01/11/2021].
285. Illumina. *How to interpret DNA strand and allele information for Infinium genotyping array data*. 2021 [cited 2021 01/11/2021]; Available from: https://emea.support.illumina.com/bulletins/2017/06/how-to-interpret-dna-strand-and-allele-information-for-infinium-.html.
286. Illumina. *Infinium Multi-Ethnic Global BeadChip*. 2016 [cited 2019 21/06/2021]; Available from: https://www.illumina.com/content/dam/illumina-

marketing/documents/products/datasheets/multi-ethnic-global-data-sheet-370-2016-001.pdf.
287. Illumina. *Infinium® HD Assay Ultra Protocol Guide*. 2009 21/06/2021].
288. Illumina. *User manual. Illumina Infinium LCG Assay Manual Protocol Experienced User Card (15023140 A)*. 2011 21/06/2021]; Available from: https://manualzz.com/doc/16830874/illumina-infinium-lcg-assay-manual-protocol-experienced-u...?__cf_chl_jschl_tk__=5532bf92d686be7814dd01976fc4955514cf07c7-1624293690-0-AYwua8M9CTPm9R1RDLsyNzVW3DQJNOYYCOw47Sb6cu7pMayxfyfWnv1hW62ElUClzpFrUkrroFBoK7QVsr_Yc3UXbL8XN_t7PtCMt9-YojYHDkSJaH3ZOgz8VEPdwm0NCAhzpcIx92teMK0_VCsQE7-PMNODVkWsNxvwHgpOcogMUZfCVi7e4uSTB9UTkAKNDW04EzbG5jOizcF7ke_-3GPNy0PKBPgFAuoRoN-uLkmnpFKE7SI4De_zT_QfIxS9C752QMr0vKfuJkInTLequbZ3-EridJ3M6swLDwfH5TXaJKJVZAmCRTVWVTXnuXBZGKEKHgRmN43yICodFECdaIUMeMseXxKToVUNdEAu7fv-ul5yyy5b6ETOf5ErmJfSMNOPQ5cPxnbF_zoRaZq20qRpIZN_b6R2YqeTWa6l4qDJCcuVoxGZ9xFBfjzAwi9kzJy6Idzp9Qx0cGUhCjWNOE32MZaGN-6_P68cU8lz5Hp_c9JtvjATTs7C2EHsBoVSIQ.
289. Illumina. *Genomestudio software*. 2011 08/03/2018]; Available from: https://emea.illumina.com/techniques/microarrays/array-data-analysis-experimental-design/genomestudio.html.
290. Illumina. *Infinium® Genotyping Data Analysis*. 2014 [cited 2019 14/01/2019]; Available from: https://www.illumina.com/Documents/products/technotes/technote_infinium_genotyping_data_analysis.pdf.
291. European Genome-Phenome Archive. *Healthy volunteer collection of European Ancestry*. 2012 [cited 2018; Available from: https://ega-archive.org/datasets/EGAD00010000144.
292. European Genome-Phenome Archive. *Healthy volunteer collection of European Ancestry*. 2014 [cited 2018; Available from: https://ega-archive.org/datasets/EGAD00010000520.
293. Illumina. *Illumina genotyping control database*. 2010 [cited 2018; Available from: https://www.illumina.com/documents/icontroldb/document_purpose.pdf.
294. Wellcome Trust. *Wellcome Trust Case Control Consortium 2*. 2008 [cited 2018; Available from: https://www.wtccc.org.uk/ccc2/.
295. Power, C. and J. Elliott, *Cohort profile: 1958 British birth cohort (National Child Development Study).* Int J Epidemiol, 2006. **35**(1): p. 34-41.
296. Anderson, C.A., et al., *Data quality control in genetic case-control association studies.* Nat Protoc, 2010. **5**(9): p. 1564-73.
297. Browning, S.R. and B.L. Browning, *Identity by Descent Between Distant Relatives: Detection and Applications.* Annual Review of Genetics, 2012. **46**(1): p. 617-633.
298. Howie, B.N., P. Donnelly, and J. Marchini. *IMPUTE2*. 2009 [cited 2021 30/06/2021]; Available from: http://mathgen.stats.ox.ac.uk/impute/impute_v2.html.
299. *Accounting for sex in the genome.* Nature Medicine, 2017. **23**(11): p. 1243-1243.

300. Edwards, A.W.F., *G. H. Hardy (1908) and Hardy–Weinberg Equilibrium.* Genetics, 2008. **179**(3): p. 1143-1150.
301. Graffelman, J. and V. Moreno, *The mid p-value in exact tests for Hardy-Weinberg equilibrium.* Statistical Applications in Genetics and Molecular Biology, 2013. **12**(4): p. 433-448.
302. Chen, B., J.W. Cole, and C. Grond-Ginsbach, *Departure from Hardy Weinberg equilibrium and genotyping error.* Frontiers in genetics, 2017. **8**: p. 167.
303. Wittke-Thompson, J.K., A. Pluzhnikov, and N.J. Cox, *Rational inferences about departures from Hardy-Weinberg equilibrium.* The American Journal of Human Genetics, 2005. **76**(6): p. 967-986.
304. Tian, C., P.K. Gregersen, and M.F. Seldin, *Accounting for ancestry: population substructure and genome-wide association studies.* Human Molecular Genetics, 2008. **17**(R2): p. R143-R150.
305. Ringnér, M., *What is principal component analysis?* Nature Biotechnology, 2008. **26**(3): p. 303-304.
306. Price, A.L., et al., *Principal components analysis corrects for stratification in genome-wide association studies.* Nature Genetics, 2006. **38**(8): p. 904-909.
307. Dufek, S., *Genome wide association study in steroid sensitive nephrotic syndrome*, in *Department of Renal Medicine*. 2019, University College London: London.
308. Li, Q. and K. Yu, *Improved correction for population stratification in genome-wide association studies by identifying hidden population structures.* Genet Epidemiol, 2008. **32**(3): p. 215-26.
309. Clarke, L., et al., *The international Genome sample resource (IGSR): A worldwide collection of genome variation incorporating the 1000 Genomes Project data.* Nucleic Acids Research, 2016. **45**(D1): p. D854-D859.
310. Devlin, B. and K. Roeder, *Genomic control for association studies.* Biometrics, 1999. **55**(4): p. 997-1004.
311. Zeng, P., et al., *Statistical analysis for genome-wide association study.* Journal of biomedical research, 2015. **29**(4): p. 285-297.
312. Chou, W.-C., et al., *A combined reference panel from the 1000 Genomes and UK10K projects improved rare variant imputation in European and Chinese samples.* Scientific Reports, 2016. **6**(1): p. 39313.
313. Belsare, S., et al., *Evaluating the quality of the 1000 genomes project data.* BMC Genomics, 2019. **20**(1): p. 620.
314. Browning, B.L., *Conform-gt*. 2016: University of Washington.
315. National Institute of Diabetes and Digestive and Kidney disease. *T1DGC HLA reference panel for inputation with SNP2HLA (T1DGC-Special)*. 2013 [cited 2021 08/07/2021]; Available from: https://repository.niddk.nih.gov/studies/t1dgc-special/?query=T1DGC%20reference%20panel.
316. Noble, J.A. and A.M. Valdes, *Genetics of the HLA region in the prediction of type 1 diabetes.* Current diabetes reports, 2011. **11**(6): p. 533-542.
317. Sperandei, S., *Understanding logistic regression analysis.* Biochemia medica, 2014. **24**(1): p. 12-18.
318. Dudbridge, F. and A. Gusnanto, *Estimation of significance thresholds for genomewide association scans.* Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society, 2008. **32**(3): p. 227-234.

319. Wang, M.H., et al., *A fast and powerful W-test for pairwise epistasis testing.* Nucleic acids research, 2016. **44**(12): p. e115-e115.
320. Sun, R., et al., *wtest: an integrated R package for genetic epistasis testing.* BMC Medical Genomics, 2019. **12**(9): p. 180.
321. Illumina. *Infinium OmniExpress-24 v1.2*. 2016 [cited 2020 07/02/2020]; Available from: https://emea.support.illumina.com/downloads/infinium-omniexpress-24-v1-2-product-files.html.
322. Dufek, S., et al., *Genetic Identification of Two Novel Loci Associated with Steroid-Sensitive Nephrotic Syndrome.* J Am Soc Nephrol, 2019. **30**(8): p. 1375-1384.
323. R Core Team, *R: A language and environment for statistical computing.* , R Foundation for Statistical Computing, Editor. 2013, R Foundation for Statistical Computing,: Vienna, Austria.
324. Wickham, H., *ggplot2: elegant graphics for data analysis*. 2016: Springer-Verlag New York.
325. Turner, S., *qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots.* The Journal of Open Source Software, 2018.
326. Millard, S., *EnvStats: An R Package for environmental Statistics*. 2013, Springer: New York.
327. Fairfax, B.P., et al., *Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression.* Science, 2014. **343**(6175): p. 1246949.
328. Fairfax, B.P., et al., *Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles.* Nat Genet, 2012. **44**(5): p. 502-10.
329. Sudlow, C., et al., *UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age.* PLoS medicine, 2015. **12**(3): p. e1001779.
330. UK Biobank. *Accessing your data*. 2020; Available from: https://biobank.ctsu.ox.ac.uk/crystal/exinfo.cgi?src=accessing_data_guide#download.
331. Band, G. and J. Marchini, *BGEN: a binary file format for imputed genotype and haplotype data.* bioRxiv, 2018: p. 308296.
332. Manichaikul, A., et al., *Robust relationship inference in genome-wide association studies.* Bioinformatics (Oxford, England), 2010. **26**(22): p. 2867-2873.
333. Bycroft, C., et al., *Genome-wide genetic data on ~500,000 UK Biobank participants.* bioRxiv, 2017: p. 166298.
334. Patterson, N., A.L. Price, and D. Reich, *Population structure and eigenanalysis.* PLoS genetics, 2006. **2**(12): p. e190.
335. Ng, P.C. and S. Henikoff, *SIFT: Predicting amino acid changes that affect protein function.* Nucleic acids research, 2003. **31**(13): p. 3812-3814.
336. Sim, N.L., et al., *SIFT web server: predicting effects of amino acid substitutions on proteins.* Nucleic Acids Res, 2012. **40**(Web Server issue): p. W452-7.
337. Kim, S., et al., *Single nucleotide polymorphisms in the phospholipase A2 receptor gene are associated with genetic susceptibility to idiopathic membranous nephropathy.* Nephron Clin Pract, 2011. **117**(3): p. c253-8.
338. OMIM. *604191 - ZINC FINGER PROTEIN 263; ZNF263*. [cited 2017 31/08/2017]; Available from: https://www.omim.org/entry/604191.

339. Landt, S.G., et al., *ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia.* Genome Res, 2012. **22**(9): p. 1813-31.
340. Bailey, T.L., et al., *MEME SUITE: tools for motif discovery and searching.* Nucleic Acids Res, 2009. **37**(Web Server issue): p. W202-8.
341. Bailey, T.L. and M. Gribskov, *Combining evidence using p-values: application to sequence homology searches.* Bioinformatics, 1998. **14**(1): p. 48-54.
342. Osada, S., et al., *DNA Binding Specificity of the CCAAT/Enhancer-binding Protein Transcription Factor Family (∗).* Journal of Biological Chemistry, 1996. **271**(7): p. 3891-3896.
343. Kel, A.E., et al., *MATCH: A tool for searching transcription factor binding sites in DNA sequences.* Nucleic Acids Res, 2003. **31**(13): p. 3576-9.
344. Agrawal, P., et al., *Genome-level identification of targets of Hox protein Ultrabithorax in Drosophila: novel mechanisms for target selection.* Sci Rep, 2011. **1**: p. 205.
345. Uniprot. *Ubx - Homeotic protein ultrabithorax - Drosophila melanogaster (Fruit fly) - Ubx gene & protein*. 2017 [cited 2017 03/08/2017]; Available from: https://www.uniprot.org/uniprot/P83949.
346. Chekmenev, D.S., C. Haid, and A.E. Kel, *P-Match: transcription factor binding site search by combining patterns and weight matrices.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W432-7.
347. Deyneko, I.V., et al., *MatrixCatch-a novel tool for the recognition of composite regulatory elements in promoters.* BMC bioinformatics, 2013. **14**(1): p. 1-10.
348. Prestridge, D.S., *SIGNAL SCAN: a computer program that scans DNA sequences for eukaryotic transcriptional elements.* Bioinformatics, 1991. **7**(2): p. 203-206.
349. Grabe, N., *AliBaba2: context specific identification of transcription factor binding sites.* In silico biology, 2002. **2**(1): p. S1-S15.
350. Inserm. *UMR_S910 - Aix Marseille Université. Human Splicing Finder - Version 3.0*. 2013 [cited 2017 13/08/2017]; Available from: http://www.umd.be/HSF3/HSF.shtml.
351. Yeo, G. and C.B. Burge, *Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals.* Journal of computational biology, 2004. **11**(2-3): p. 377-394.
352. Bycroft, C., et al., *The UK Biobank resource with deep phenotyping and genomic data.* Nature, 2018. **562**(7726): p. 203-209.
353. Privé, F., et al., *Efficient toolkit implementing best practices for principal component analysis of population genetic data.* bioRxiv, 2019: p. 841452.
354. Altman, N. and M. Krzywinski, *Points of Significance: Association, correlation and causation.* Nature methods, 2015. **12**(10).
355. Chinery, R., et al., *Antioxidant-induced nuclear translocation of CCAAT/enhancer-binding protein β: a critical role for protein kinase A-mediated phosphorylation of Ser299.* Journal of Biological Chemistry, 1997. **272**(48): p. 30356-30361.
356. Roy, S.K., et al., *MEKK1 plays a critical role in activating the transcription factor C/EBP-beta-dependent gene expression in response to IFN-gamma.* Proc Natl Acad Sci U S A, 2002. **99**(12): p. 7945-50.
357. Pless, O., et al., *G9a-mediated lysine methylation alters the function of CCAAT/enhancer-binding protein-beta.* J Biol Chem, 2008. **283**(39): p. 26357-63.

358. Jamaluddin, M., et al., *Inducible translational regulation of the NF-IL6 transcription factor by respiratory syncytial virus infection in pulmonary epithelial cells.* Journal of Virology, 1996. **70**(3): p. 1554-1563.
359. Nolin, J.D., et al., *Identification of epithelial phospholipase A2 receptor 1 as a potential target in asthma.* American journal of respiratory cell and molecular biology, 2016. **55**(6): p. 825-836.
360. Assandri, R., et al., *Anti-Phospholipase A2 receptor antibodies in membranous nephropathy: from bench to patient.* Journal of Nephrology & Therapeutics, 2014. **4**(2): p. 1000155.
361. GeneCards. *CEBPB Gene - CCAAT Enhancer Binding Protein Beta*. 2021 [cited 2021 14/12/2021]; Available from: https://www.genecards.org/cgi-bin/carddisp.pl?gene=CEBPB.
362. Tiller, G.E. and V.A. McKusick. *189965. CCAAT/ENHANCER-BINDING PROTEIN, BETA; CEBPB*. 2013 [cited 02/08/2017; Available from: https://www.omim.org/entry/189965.
363. Millward, C.A., et al., *Mice with a deletion in the gene for CCAAT/enhancer-binding protein beta are protected against diet-induced obesity.* Diabetes, 2007. **56**(1): p. 161-7.
364. Brommage, R., et al., *High-throughput screening of mouse gene knockouts identifies established and novel skeletal phenotypes.* Bone Res, 2014. **2**: p. 14034.
365. Xu, L., et al., *Immune and inflammatory mechanisms. MP362: Expression and role of CEBP subtypes in chronic kidney disease.* Nephrology Dialysis Transplantation, 2013. **28**(suppl_1): p. i406-i414.
366. Qin, W., et al., *Anti-phospholipase A2 receptor antibody in membranous nephropathy.* J Am Soc Nephrol, 2011. **22**(6): p. 1137-43.
367. Ifuku, M., et al., *Various roles of Th cytokine mRNA expression in different forms of glomerulonephritis.* Am J Nephrol, 2013. **38**(2): p. 115-23.
368. Jefferson, J.A., J.W. Pippin, and S.J. Shankland, *Experimental Models of Membranous Nephropathy.* Drug Discov Today Dis Models, 2010. **7**(1-2): p. 27-33.
369. Harbison, C.T., et al., *Transcriptional regulatory code of a eukaryotic genome.* Nature, 2004. **431**(7004): p. 99-104.
370. Omnibus, G.E. *Sample GSM1010889*. 2012 [cited 2021 14/12/2021]; Available from: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1010889.
371. UniProt. *Search results 'transcription factor' AND organism:"Homo sapiens (Human) [9606]" in UniProtKB*. 2021 [cited 2021 12/08/2021]; Available from: https://www.uniprot.org/uniprot/?query=transcription+factors+organism%3A%22Homo+sapiens+%28Human%29+%5B9606%5D%22&sort=score.
372. Gene Ontology 1.8. *AmiGO 2.5.15 Search engine*. 2021 [cited 2021; 12/08/2021]. Available from: http://amigo.geneontology.org/amigo/medial_search?q=homo%2Bsapiens%2Btranscription%2Bfactor.
373. Wilson, D., et al., *DBD--taxonomically broad transcription factor predictions: new content and functionality.* Nucleic Acids Res, 2008. **36**(Database issue): p. D88-92.
374. Kummerfeld, S.K. and S.A. Teichmann, *DBD: a transcription factor prediction database.* Nucleic Acids Res, 2006. **34**(Database issue): p. D74-81.

375. Consortium, E.P., *A user's guide to the encyclopedia of DNA elements (ENCODE).* PLoS Biol, 2011. **9**(4): p. e1001046.
376. Uniprot. *UniProtKB - P01100 (FOS_HUMAN)*. 2021 [cited 2021 12/08/2021]; Available from: https://www.uniprot.org/uniprot/P01100.
377. Torban, E. and P. Goodyer, *What PAX genes do in the kidney.* Nephron Experimental Nephrology, 1998. **6**(1): p. 7-11.
378. National centre for biotechnology information. *PAX6 paired box 6 [Homo sapiens (human)] - Gene*. 2021 [cited 2021 12/08/2021]; Available from: https://www.ncbi.nlm.nih.gov/gene/5080.
379. Butta, N., et al., *Role of transcription factor Sp1 and CpG methylation on the regulation of the human podocalyxin gene promoter.* BMC Mol Biol, 2006. **7**: p. 17.
380. Meisler, M.H., *Evolutionarily conserved noncoding DNA in the human genome: how much and what for?* Genome Res, 2001. **11**(10): p. 1617-8.
381. Friedli, M., et al., *A Systematic Enhancer Screen Using Lentivector Transgenesis Identifies Conserved and Non-Conserved Functional Elements at the Olig1 and Olig2 Locus.* PLOS ONE, 2011. **5**(12): p. e15741.
382. Chatterjee, S., G. Bourque, and T. Lufkin, *Conserved and non-conserved enhancers direct tissue specific transcription in ancient germ layer specific developmental control genes.* BMC Developmental Biology, 2011. **11**(1): p. 63.
383. Geertz, M. and S.J. Maerkl, *Experimental strategies for studying transcription factor-DNA binding specificities.* Brief Funct Genomics, 2010. **9**(5-6): p. 362-73.
384. Dhaouadi, T., et al., *PLA2R antibody, PLA2R rs4664308 polymorphism and PLA2R mRNA levels in Tunisian patients with primary membranous nephritis.* PLOS ONE, 2020. **15**(10): p. e0240025.
385. Abramovitz, M., et al., *Optimization of RNA extraction from FFPE tissues for expression profiling in the DASL assay.* Biotechniques, 2008. **44**(3): p. 417-23.
386. Wang, F., et al., *RNAscope: a novel in situ RNA analysis platform for formalin-fixed, paraffin-embedded tissues.* J Mol Diagn, 2012. **14**(1): p. 22-9.
387. Nucleics. *Failed DNA Sequencing Reactions*. 2021 [cited 2021 12/08/2021]; Available from: https://www.nucleics.com/DNA_sequencing_support/DNA-sequencing-failed-reaction.html.
388. Thermo Fisher Scientific. *ExoSAP-ITTM PCR Product Cleanup Reagent*. 2018 [cited 2021 12/08/2021]; Available from: https://www.thermofisher.com/order/catalog/product/78200.200.UL.
389. Oxford Nanopore Technologies. *DNA: nanopore sequencing*. 2018 [cited 2021 12/08/2021]; Available from: https://nanoporetech.com/applications/dnananopore-sequencing.
390. Heyn, P., et al., *Introns and gene expression: cellular constraints, transcriptional regulation, and evolutionary consequences.* Bioessays, 2015. **37**(2): p. 148-54.
391. Brenet, F., et al., *DNA methylation of the first exon is tightly linked to transcriptional silencing.* PLoS One, 2011. **6**(1): p. e14524.
392. Brandt, D.Y., et al., *Mapping Bias Overestimates Reference Allele Frequencies at the HLA Genes in the 1000 Genomes Project Phase I Data.* G3 (Bethesda), 2015. **5**(5): p. 931-41.

393. Kichaev, G., et al., *Integrating Functional Data to Prioritize Causal Variants in Statistical Fine-Mapping Studies.* PLOS Genetics, 2014. **10**(10): p. e1004722.
394. Li, Y.R. and B.J. Keating, *Trans-ethnic genome-wide association studies: advantages and challenges of mapping in diverse populations.* Genome Medicine, 2014. **6**(10): p. 91.
395. *Genetics for all.* Nature Genetics, 2019. **51**(4): p. 579-579.
396. Skol, A.D., et al., *Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies.* Nature genetics, 2006. **38**(2): p. 209-213.
397. Johnson, L. *GAS Power Calculator*. 2017 [cited 31/10/2019; Available from: http://csg.sph.umich.edu/abecasis/cats/gas_power_calculator/index.html.
398. Verlouw, J.A.M., et al., *A comparison of genotyping arrays.* European Journal of Human Genetics, 2021. **29**(11): p. 1611-1624.
399. Verma, S.S., et al., *Imputation and quality control steps for combining multiple genome-wide datasets.* Frontiers in Genetics, 2014. **5**(370).
400. Tao. *Imputation on two genotyping datasets*. 2016; Available from: https://www.biostars.org/p/221993/.
401. Cordell, H. *Imputation and haplotyping for genome-wide association analysis*. in *Genetic Analysis of Mendelian and Complex Disorders*. 2021. Virtual.
402. Mitt, M., et al., *Improved imputation accuracy of rare and low-frequency variants using population-specific high-coverage WGS-based imputation reference panel.* European Journal of Human Genetics, 2017. **25**(7): p. 869-876.
403. Yoo, S.-K., et al., *NARD: whole-genome reference panel of 1779 Northeast Asians improves imputation accuracy of rare and low-frequency variants.* Genome Medicine, 2019. **11**(1): p. 64.
404. Schork, N.J., et al., *Common vs. rare allele hypotheses for complex diseases.* Current opinion in genetics & development, 2009. **19**(3): p. 212-219.
405. El-Fishawy, P., *Common Disease-Common Variant Hypothesis*, in *Encyclopedia of Autism Spectrum Disorders*, F.R. Volkmar, Editor. 2013, Springer New York: New York, NY. p. 719-720.
406. Ooi, J.D., et al., *Dominant protection from HLA-linked autoimmunity by antigen-specific regulatory T cells.* Nature, 2017. **545**(7653): p. 243-247.
407. Lyons, P.A., et al., *Genetically distinct subsets within ANCA-associated vasculitis.* N Engl J Med, 2012. **367**(3): p. 214-23.
408. Dai, H., H. Zhang, and Y. He, *Diagnostic accuracy of PLA2R autoantibodies and glomerular staining for the differentiation of idiopathic and secondary membranous nephropathy: an updated meta-analysis.* Sci Rep, 2015. **5**: p. 8803.
409. Pourcine, F., et al., *Prognostic value of PLA2R autoimmunity detected by measurement of anti-PLA2R antibodies combined with detection of PLA2R antigen in membranous nephropathy: A single-centre study over 14 years.* PLoS One, 2017. **12**(3): p. e0173201.
410. Francis, J.M., L.H. Beck, Jr., and D.J. Salant, *Membranous Nephropathy: A Journey From Bench to Bedside.* Am J Kidney Dis, 2016. **68**(1): p. 138-47.
411. Timmermans, S.A., et al., *Evaluation of anti-PLA2R1 as measured by a novel ELISA in patients with idiopathic membranous nephropathy: a cohort study.* Am J Clin Pathol, 2014. **142**(1): p. 29-34.
412. Gupta, S., et al., *A genetic risk score distinguishes different types of autoantibody mediated membranous nephropathy*. 2021, UCL: KI Reports.

413. Stelzer, G., et al., *The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses.* Current Protocols in Bioinformatics, 2016. **54**(1): p. 1.30.1-1.30.33.
414. Weitzmann Institute of Science and LifeMap Sciences. *GeneCards®: The Human Gene Database*. GeneCards Version 5.6 2021 [cited 2021 16/11/2021]; Available from: https://www.genecards.org/.
415. Consortium, T.U., *UniProt: the universal protein knowledgebase in 2021.* Nucleic Acids Research, 2020. **49**(D1): p. D480-D489.
416. Debiec, H., et al., *Transethnic, Genome-Wide Analysis Reveals Immune-Related Risk Alleles and Phenotypic Correlates in Pediatric Steroid-Sensitive Nephrotic Syndrome.* J Am Soc Nephrol, 2018. **29**(7): p. 2000-2013.
417. Dyck, R., C. Bohm, and H. Klomp, *Increased frequency of HLA A2/DR4 and A2/DR8 haplotypes in young saskatchewan aboriginal people with diabetic end-stage renal disease.* American journal of nephrology, 2003. **23**(3): p. 178-185.
418. Liu, D., et al., *Gene polymorphism and risk of idiopathic membranous nephropathy.* Life Sciences, 2019. **229**: p. 124-131.
419. Kumar, V., et al., *Primary membranous nephropathy in adolescence: A prospective study.* Nephrology, 2017. **22**(9): p. 678-683.
420. European Molecular Biology Laboratory - European Bioinformatics Institute. *The Polygenic Score Catalog*. 2021 [cited 2021 16/08/2021]; Available from: https://www.pgscatalog.org/.
421. Natarajan, P., et al., *Polygenic risk score identifies subgroup with higher burden of atherosclerosis and greater relative benefit from statin therapy in the primary prevention setting.* Circulation, 2017. **135**(22): p. 2091-2101.
422. Kaga, H., et al., *Comparison of measurements of anti-PLA2R antibodies in Japanese patients with membranous nephropathy using in-house and commercial ELISA.* Clinical and experimental nephrology, 2019. **23**(4): p. 465-473.
423. Janssens, A.C.J.W., *Proprietary Algorithms for Polygenic Risk: Protecting Scientific Innovation or Hiding the Lack of It?* Genes, 2019. **10**(6): p. 448.
424. Cremoni, M., et al., *Th17-Immune response in patients with membranous nephropathy is associated with thrombosis and relapses.* Frontiers in Immunology, 2020. **11**: p. 3073.
425. Liu, W., et al., *Idiopathic Membranous Nephropathy: Glomerular Pathological Pattern Caused by Extrarenal Immunity Activity.* Frontiers in Immunology, 2020. **11**(1846).
426. Couser, W.G. and R.J. Johnson, *The etiology of glomerulonephritis: roles of infection and autoimmunity.* Kidney Int, 2014. **86**(5): p. 905-14.
427. Gola, D., et al., *Population Bias in Polygenic Risk Prediction Models for Coronary Artery Disease.* Circulation: Genomic and Precision Medicine, 2020. **13**(6): p. e002932.
428. Zhong, H. and R.L. Prentice, *Bias-reduced estimators and confidence intervals for odds ratios in genome-wide association studies.* Biostatistics, 2008. **9**(4): p. 621-634.
429. Sharp, S.A., et al., *Development and Standardization of an Improved Type 1 Diabetes Genetic Risk Score for Use in Newborn Screening and Incident Diagnosis.* Diabetes care, 2019. **42**(2): p. 200-207.
430. Rose, G., *Sick individuals and sick populations. 1985.* Bull World Health Organ, 2001. **79**(10): p. 990-6.

431. Hamilton, P., et al., *The investigative burden of membranous nephropathy in the UK.* Clinical Kidney Journal, 2019. **13**(1): p. 27-34.
432. Stubbs, M.J., *Characterisation of immune thrombotic thrombocytopenic purpura utilising genome-wide association study*, in *Department of Renal Medicine*. 2020, UCL: London.
433. Bansal, V. and O. Libiger, *Fast individual ancestry inference from DNA sequence data leveraging allele frequencies for multiple populations.* BMC Bioinformatics, 2015. **16**(1): p. 4.
434. Jin, Y., et al., *GRAF-pop: A Fast Distance-Based Method To Infer Subject Ancestry from Multiple Genotype Datasets Without Principal Components Analysis.* G3: Genes|Genomes|Genetics, 2019. **9**(8): p. 2447-2461.

10. Publications (during PhD)

Gupta S, Downie ML, Cheshire C, Dufek-Kamperis S, Levine AP, Brenchley P, Hoxha E, Stahl RAK, Ashman N, Pepper RJ, Mason S, Norman J, Bockenhauer D, Stanescu HC, Kleta R, Gale DP. **(2022)** A genetic risk score distinguishes different types of autoantibody mediated membranous nephropathy. In progress.
Downie ML, **Gupta S**, Chan MMY, Sadeghi-Alavijeh O, Cao J, Parekh R, Diz CB, Bierzynska A, Genomics England Research Consortium, NIHR BioResource, Levine AP, Pepper RJ, Stanescu H, Saleem MA, Kleta R, Bockenhauer D, Koziell AB, Gale DP. **(2022)** Shared genetic risk across different presentations of gene test-negative idiopathic nephrotic syndrome. *Kidney International.* In progress.
Hamilton P, Blaikie K, Roberts SA, Gittins M, Downie ML, **Gupta S**, Voinescu C, Kanigicherla D, Stanescu HC, Kleta R, Brenchley P. **(2022)** Membranous Nephropathy in the UK Biobank. *Journal of American Society of Nephrology.* In progress.
Downie ML, **Gupta S**, Tekman M, Cheshire C, Arora S, Licht C, Robinson L, Munoz M, Aris AM, Al-Attrach I, Brenchley PE, Gale DP, Stanescu H, Bockenhauer D, Kleta R. **(2021)** Identification of a locus on the X-chromosome linked to familial membranous nephropathy. *Kidney International Reports.* 6(6):1669-1676
Xie J, Liu L… **Gupta S**… Kleta R, Chan N, Kiryluk K. **(2020)** The genetic architecture of membranous nephropathy and its potential to improve non-invasive diagnosis. *Nature Communications*. 11:1600
Dufek S, Cheshire C, Levine AP, Trompeter RS, Issler N, Stubbs M, Mozere M, **Gupta S**, … Gale D, Stanescu HC, Kleta R, Bockenhauer D. **(2019)** Genetic identification of CALHM6 as novel locus for steroid-sensitive nephrotic syndrome. *Journal of American Society of Nephrology*. 30(8):1375-1384

Mathew D, **Gupta S,** Ashman N. **(2021)** A case report of breast cancer and membranous nephropathy with positive anti-phospholipase A2 receptor antibodies. *BMC Nephrology.* 22:324
Nikolopoulou A, **Gupta S,** Tulley C, Stuart J, Pepper RJ, Ashman N, Griffith M. **(2021)** Rituximab dosing in membranous nephropathy may be suboptimal. *Poster presentation, UK Kidney Week, Renal Association*
Gupta S, Nikolopoulou A, Connolly J, Oates T, Pepper RJ, Griffith M, Ashman N. **(2019)** The London Membranous Network. *Poster presentation, UK Kidney Week, Renal Association*
Gupta S, Pepper RJ, Ashman N, Walsh SB. **(2018)** Nephrotic Syndrome: Oedema Formation and Its Treatment With Diuretics. *Frontiers in Physiology* 9:1868
Gupta S, Kottgen A, Hoxha E, Brenchley P, Bockenhauer D, Stanescu HC, Kleta R. **(2017)** Genetics of membranous nephropathy. *Nephrology, Dialysis and Transplantation* 1-9
Gupta S, Connolly J, Pepper RJ, Walsh SB, Yaqoob MM, Kleta R, Ashman N. **(2017)** Membranous Nephropathy: A retrospective observational study of membranous nephropathy in North East and Central London. *BMC Nephrology* 18:201
Wills M, **Gupta S**, Gage A, Ashman N, Forbes S. **(2020)** The direct use of oral anti-coagulants in membranous nephropathy. *Nephrology Dialysis Transplantation.* 35(3):P0467
Oates T, Gleeson S, **Gupta S**, Nikolopoulou L, Pepper RJ, Griffith M, Ashman N. **(2020)** Strategies for reduction of cardiovascular risk: effect of time and different treatments on lipids in membranous GN. *Oral abstract presentation, UK Kidney Week, Renal Association.* P265

11. Appendix

The following are the accepted versions of the publications that arose during my PhD. There are a few that are currently under review and these have not been included. The order matches that as in section 10.

KI REPORTS
KIReports.org
CLINICAL RESEARCH

Identification of a Locus on the X Chromosome Linked to Familial Membranous Nephropathy

Mallory L. Downie[1,2], Sanjana Gupta[1], Mehmet C. Tekman[3], Chris Cheshire[1], Steven Arora[4], Christoph Licht[5], Lisa A. Robinson[5], Marina Munoz[6], Alvaro Madrid Aris[7], Ibrahim Al Attrach[8], Paul E. Brenchley[9], Daniel P. Gale[1], Horia Stanescu[1], Detlef Bockenhauer[1,2] and Robert Kleta[1,2]

[1]Department of Renal Medicine, University College London, London, UK; [2]Paediatric Nephrology, Great Ormond Street Hospital for Children National Health Service Foundation Trust, London, UK; [3]Department of Computer Science, University of Freiburg, Freiburg, Germany; [4]Department of Paediatrics, Division of Nephrology, McMaster Children's Hospital, Hamilton, Ontario, Canada; [5]Division of Nephrology, The Hospital for Sick Children, Toronto, Ontario, Canada; [6]Department of Paediatric Nephrology, University Hospital Vall d'Hebron, Barcelona, Spain; [7]Department of Nephrology, Hospital Sant Joan de Deu, Barcelona, Spain; [8]Department of Pediatrics, Tawam Hospital, Al Ain, United Arab Emirates; and [9]Division of Cardiovascular Sciences, University of Manchester, Manchester, UK

Introduction: Membranous nephropathy (MN) is the most common cause of nephrotic syndrome (NS) in adults and is a leading cause of end-stage renal disease due to glomerulonephritis. Primary MN has a strong male predominance, accounting for approximately 65% of cases; yet, currently associated genetic loci are all located on autosomes. Previous reports of familial MN have suggested the existence of a potential X-linked susceptibility locus. Identification of such risk locus may provide clues to the etiology of MN.

Methods: We identified 3 families with 8 members affected by primary MN. Genotyping was performed using single-nucleotide polymorphism microarrays, and serum was sent for anti-phospholipase A2 receptor (PLA2R) antibody testing. All affected members were male and connected through the maternal line, consistent with X-linked inheritance. Genome-wide multipoint parametric linkage analysis using a model of X-linked recessive inheritance was conducted, and genetic risk scores (GRSs) based on known MN-associated variants were determined.

Results: Anti-PLA2R testing was negative in all affected family members. Linkage analysis revealed a significant logarithm of the odds score (3.260) on the short arm of the X chromosome at a locus of approximately 11 megabases (Mb). Haplotype reconstruction further uncovered a shared haplotype spanning 2 Mb present in all affected individuals from the 3 families. GRSs in familial MN were significantly lower than in anti-PLA2R–associated MN and were not different from controls.

Conclusions: Our study identifies linkage of familial membranous nephropathy to chromosome Xp11.3-11.22. Family members affected with MN have a significantly lower GRS than individuals with anti-PLA2R–associated MN, suggesting that X-linked familial MN represents a separate etiologic entity.

Kidney Int Rep (2021) **6,** 1669–1676; https://doi.org/10.1016/j.ekir.2021.02.025
KEYWORDS: genetic risk score; glomerulonephritis; linkage analysis; LOD score; membranous nephropathy; X-linked

MN is the most common cause of NS in adults and is a leading cause of end-stage kidney disease (ESKD) due to glomerular disease.[1] Although common in adults, MN is uncommon in children and usually accounts for less than 5% of pediatric patients undergoing biopsy for NS.[2] Across ages, MN demonstrates a 2:1 male predominance and is most often a sporadic disease.[3] Rare examples of familial MN have also been reported, usually presenting in siblings.[4] Although MN can occur in the context of systemic disease (secondary), primary (or idiopathic) MN accounts for 80% of cases in adults, of which approximately one-third progress to ESKD within 5 to 15years.[1,3,5]

Importantly, 85% of individuals with primary MN have IgG4 autoantibodies against the podocyte membrane antigen PLA2R, meaning that treatment tends to

Correspondence: Mallory L. Downie, UCL Department of Renal Medicine, 1st Floor, Royal Free Hospital, Rowland Hill Street, London NW3 2PF, UK. E-mail: mallory.downie.18@ucl.ac.uk
Received 16 October 2020; revised 6 February 2021; accepted 6 February 2021; published online 3 March 2021

involve immunosuppressive therapy.[3] There remains a subset of patients with primary MN who have no identified autoantibodies and indeed have variable response to immunosuppression.[6] The etiology in this subset of patients is not yet understood, and genetic studies could provide important clues about disease mechanisms, especially in the context of familial clustering.

Although MN has a strong male predominance, currently associated alleles are all located on autosomes.[7] Genome-wide association studies implicate risk alleles in both *HLA-DQA1* and *PLA2R* genes, which contribute the highest proportion of disease risk, and in newly identified loci encoding *NFKB1* and *IRF4*, contributing a smaller proportion.[7–10] Identification of a risk locus (or loci) on the X chromosome could help explain why males are predominantly affected.

Our study investigates 3 families with idiopathic MN and negative anti-PLA2R antibodies with pedigrees suggestive of X-linked inheritance. We sought (i) to determine whether there is a risk locus on the X chromosome in these families, and (ii) to determine whether the known *HLA-DQA1*– and *PLA2R*-associated risk alleles contribute to their genetic risk.

METHODS

The study recruited 3 families of European ethnicity with 8 members affected by biopsy specimen-proven idiopathic MN. Affected members had serum tested for anti-PLA2R antibodies using the enzyme-linked immunosorbent assay.[11] Clinical features, such as age at presentation, response to immunosuppressive therapy, progression to renal failure, and renal transplant status, were also obtained from each individual's home institution, if available.

DNA was isolated from the 8 affected and 18 apparently unaffected family members from whole blood using standard procedures. Family 1 was genotyped *via* Omni-X-24 BeadChip (Illumina, San Diego, CA), with a total of 741,000 markers, and families 2 and 3 were genotyped *via* an Infinium Multi-Ethnic Global BeadChip (Illumina), with a total of 1,779,819 markers. Genotype files then underwent quality control checks as described previously.[12] Multipoint parametric linkage analysis, performed for families 1 to 3 using a model of X-linked recessive inheritance, was conducted in both Allegro (deCODE Genetics, Reykjavik, Iceland) and Merlin (University of Michigan, Ann Arbor, MI).[13,14] Alohomora (Max Delbruck Center (MDC) for Molecular Medicine Berlin-Buch, Germany) was used to generate input files for both linkage programs, and linkage output files were visualized using R 3.2.0 software (R Foundation for Statistical Computing, Vienna, Austria).[15] Haplotype reconstruction was performed and visualized in HaploForge (Free Software Foundation, Boston, MA),[16] with input files generated by Allegro. X-chromosomal regions were considered significant for linkage if the logarithm of the odds (LOD) score was >2.475 due to the lower number of recombination events on gonosomes.[17]

GRSs in our families with familial MN were calculated using the odds ratios at each autosomal risk loci, *HLA-DQA1* and *PLA2R*, determined from an independent historical genome-wide association studies analysis.[7] The GRS was computed by the sum of the natural logarithm of the odds ratio at each autosomal risk SNP multiplied by the number of risk alleles (0, 1, or 2), divided by the number of possible alleles.[18] The GRS was calculated for study individuals affected with familial MN (n = 8), unaffected (n = 18), and combined (n = 25). Scores were then compared against a European adult cohort of anti-PLA2R–positive MN patients (n = 410) and healthy European controls (n = 5642). Results were statistically compared using the χ^2 test with the Bonferroni correction for multiple comparisons. Statistical analysis and data visualization was performed in R 3.2.0 software.

RESULTS

Family Pedigrees

Pedigrees for each family are displayed in Figure 1. Pedigree analysis showed a pattern consistent with X-linked recessive inheritance in all families.

Clinical Details for Affected Family Members

All affected individuals included in the study had biopsy specimen-proven idiopathic MN and all had negative serologic test results for anti-PLA2R antibody. Histopathology images of family 1 have been published previously[4]; histopathology images for families 2 and 3 were unavailable. Histopathology features observed in the 3 families are detailed in Supplementary Table S1. No other autoimmune diseases were reported in any family member included in this study. Please refer to Figure 1 for pedigree position of each individual (identification numbers in bold) outlined in the family descriptions, below.

Family 1

Details of this British family have been reported previously.[4] Briefly:

61: Individual presented at age 3 years with NS, microscopic hematuria, and hypertension, which were initially responsive to combination therapy of corticosteroids and cyclophosphamide. Although this immunosuppression put him into remission at first, he went on to develop a relapsing course and eventually

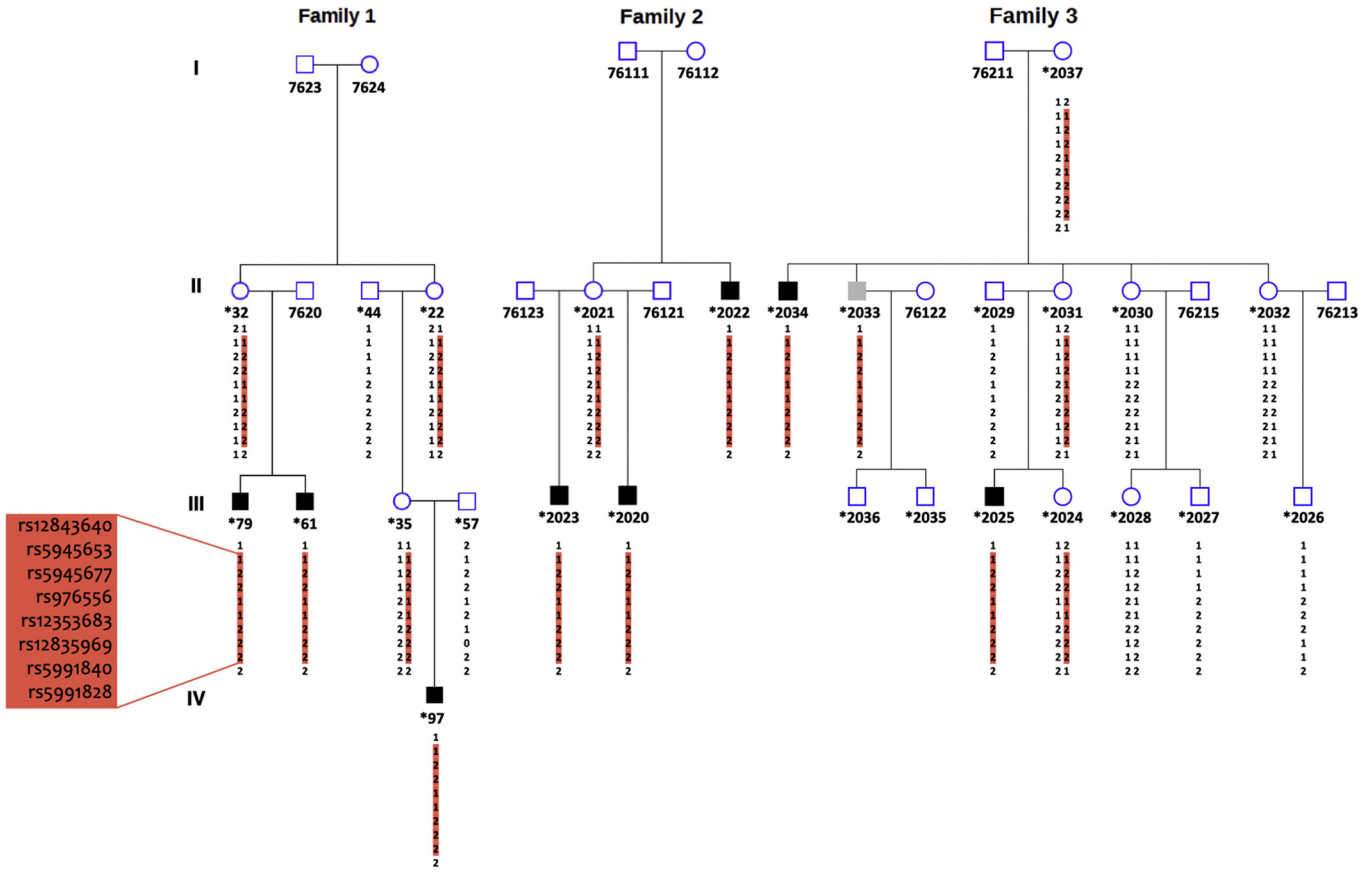


Figure 1. Pedigrees and haplotypes of families 1, 2, and 3 with familial membranous nephropathy. Squares indicate males and circles indicate females. A black symbol indicates that the individual is affected, a white symbol indicates the individual is unaffected, and a grey symbol indicates that the individual's affectation status is unknown. Asterisks indicate individuals who were genotyped and included in the study. Red boxes indicate the shared haplotype (rs12843640-rs5991828). Pedigree analysis in all 3 families showed a pattern consistent with X-linked recessive inheritance (i.e., only males are affected and inheritance is *via* the maternal line with no male-to-male transmission).

develop ESKD. He received a renal transplant at age 23 years and has not had subsequent recurrence of disease.

79: Individual presented at age 10 years with NS that was unresponsive to corticosteroids and a trial of azathioprine. He also had significant hypertension that led to a hypertensive crisis, seizures, and cerebral infarction, leaving him with permanent neurologic deficits. A spontaneous remission of NS occurred after 1 year, but it did eventually relapse. He was treated with cyclophosphamide and had some improvement. Like his brother, however, he went on to develop a relapsing disease course accompanied by declining glomerular filtration rate. At the last follow-up (age 31 years), he had chronic kidney disease stage 4.

97: Individual presented at age 1 year with NS, hematuria, and hypertension, which were unresponsive to steroids. He has had a relapsing disease course, with relapses occurring approximately every 3 months. At the last follow-up (age 16 years), he was in chronic kidney disease stage 3 and in partial remission and was being treated with mycophenolate mofetil, cyclosporin, and an angiotensin receptor blocker.

Family 2

This family resides in Canada.

2020: Individual presented at age 11 years with NS that eventually progressed to ESKD treated with a renal transplant. He then developed recurrence of disease post-transplant. Whether he had response to immunosuppression is unknown.

2023: Individual presented at age 6 years with NS that also progressed to ESKD, and like his half-brother, he developed disease recurrence after renal transplant.

2022: Individual presented in teenage period with kidney disease that progressed to ESKD. Whether he had features of NS or response to immunosuppression is unknown. He underwent 3 kidney transplants, all of which failed due to recurrence of disease.

Family 3

This family resides in Spain.

2025: Individual presented at age 5 years with NS that was treated with cyclosporine to induce full remission.

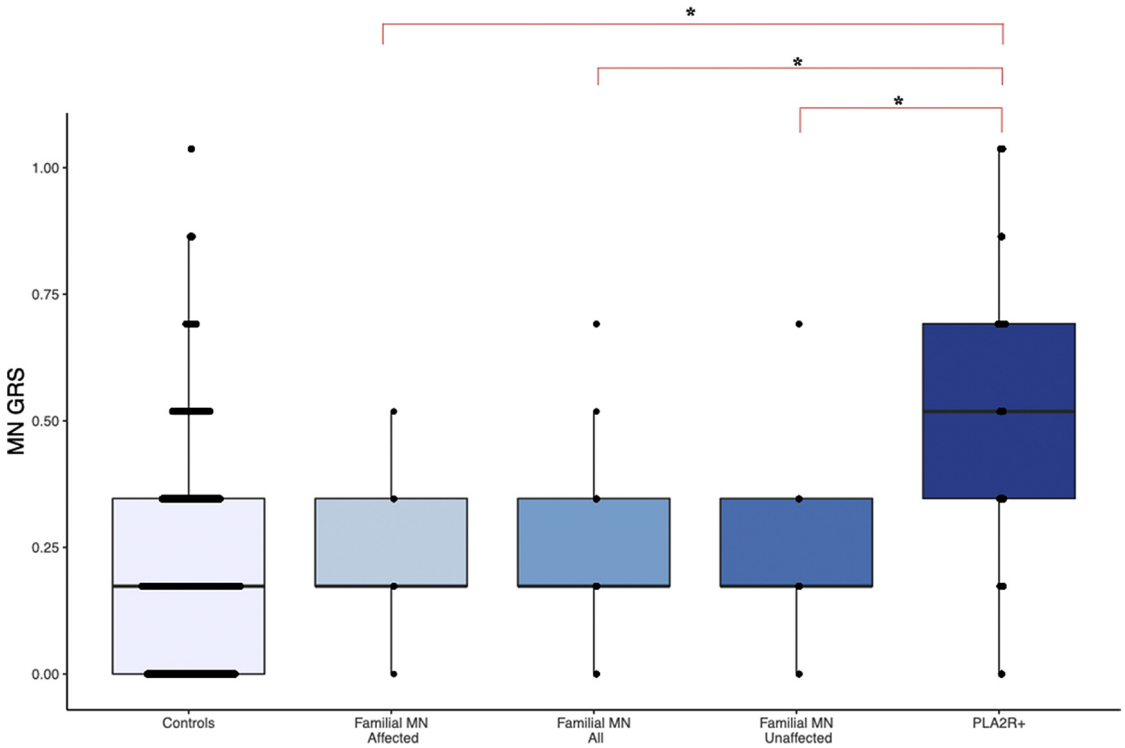


Figure 2. Box and whisker plot shows genetic risk scores (GRSs) in familial membranous nephropathy (MN). Median values (line inside the box) for each group with upper and lower quartiles (top and bottom) are represented by boxes, with whiskers delineating variability outside quartiles. Outliers are plotted as individual point beyond whisker limits. Asterisks indicate $P < 0.05$ using the χ^2 test with the Bonferroni correction.

2033: Individual was assessed at age 50 years. At that time, he had no clinical evidence of kidney disease. This family was lost to follow-up, and whether proteinuria has since developed is unknown.

2034: Individual is currently age 55 years and has a phenotype of NS. He presented with symptoms in adulthood, and whether he had any response to immunosuppression is unknown.

Genetic Risk Scores

GRSs in familial MN, calculated using risk estimates at *HLA-DQA1* and *PLA2R* loci, were found to be significantly lower than in individuals with MN associated with anti-PLA2R antibodies. GRSs in familial MN were not significantly different than controls (see Figure 2).

Linkage Analysis

Multipoint parametric linkage analysis for X-linked recessive inheritance in the 3 families initially revealed an 11-Mb region of linkage on the X chromosome. This region had a LOD score of 3.260 and had flanking markers of rs12014680 and rs2360739 (see Figure 3).

Haplotype reconstruction confirmed that the affected individuals within each family shared a haplotype that was also present as 1 allele in the unaffected "carrier" mothers. In families 1 and 2, this haplotype was not present in any other individuals; however, 1 adult man (aged 50 years) with unknown affectation status in family 3 shared the same haplotype across the linked region as the affected male individuals (see Figure 1 and Supplementary Figure S1). Because disease onset was observed in this family beyond the pediatric period (>18 years), we designated this individual as affectation status unknown so his data do not contribute to the LOD score.

Haplotype reconstruction showed flanking markers rs3027452 and rs2360739. A shared haplotype (i.e., identical alleles at all 8 markers) was further identified between all individuals carrying a risk allele, spanning a narrower region of 2 Mb (rs12843640–rs5991828), suggesting a shared distant common ancestor (founder effect).

This 11-Mb linked region on the X chromosome mapped to Xp11.3-p11.22, which contained 167 unique genes based on the Human Genome Organisation (HUGO) Official Gene Symbol listed in University of California Santa Cruz (UCSC) Genome Browser.[19] They represent a wide range of functionalities, including many ubiquitous proteins. Restricting the linked area to the 2-Mb region of the shared haplotype (Xp11.22) limited the list to 70 unique genes (see Table 1).

DISCUSSION

In this study, we present 3 families affected by renal biopsy specimen-proven idiopathic MN, mostly

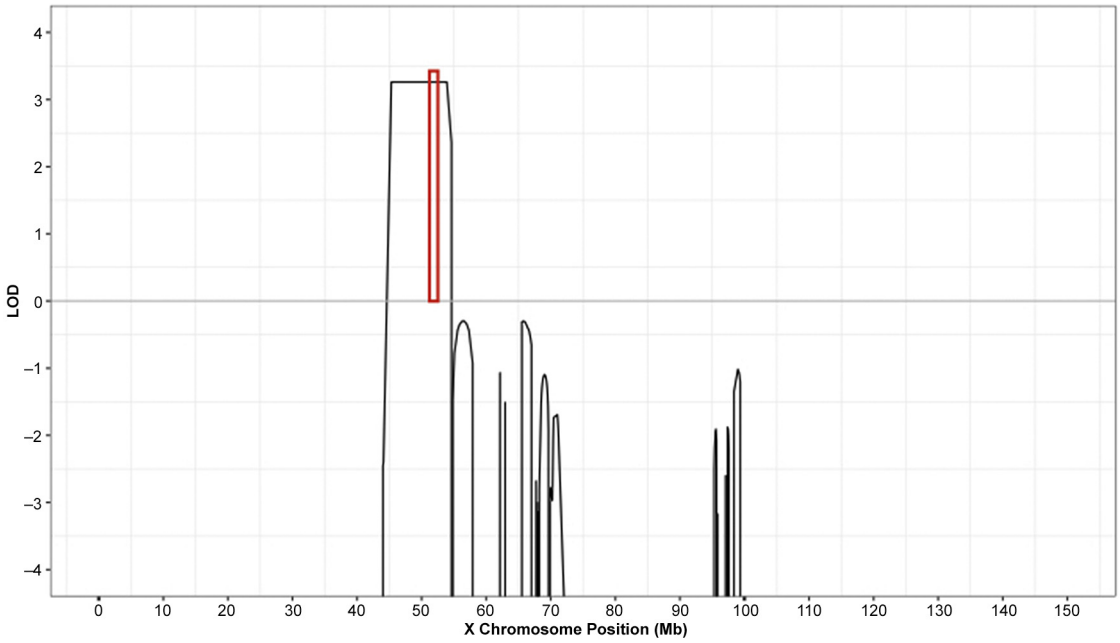


Figure 3. Multipoint parametric linkage analysis on chromosome X for families 1, 2, and 3. The *y* axis shows the logarithm of the odds (LOD) score, and the *x* axis gives the genomic position in megabases (Mb). Note significant linkage of 3.260 in the region of 43 to 54 Mb (reference genome: GRCh37). The red box indicates the area of shared identical haplotype in all 3 families (51–53 Mb).

presenting in childhood. All affected individuals were males connected through the maternal line, suggesting X-linked inheritance. By calculating GRSs in these families using risk allele counts at known autosomal risk loci, *HLA-DQA1* and *PLA2R*, we observed that the GRS was lower in familial MN compared with anti-PLA2R antibody-associated disease. Combined with the finding that all affected individuals also had negative serologic testing for anti-PLA2R antibodies, this suggested that the observed familial clustering was unlikely to be attributable to aggregation of known genetic risk factors and coincidental occurrence of the most prevalent cause of disease. In addition, although all of the families were of similar European ethnic background, they were recruited from 3 countries, and within each family, individuals from at least 2 different households were affected. These details imply that shared environmental exposures are unlikely to explain the observed familial clustering of disease.

Using X-linked recessive multipoint linkage analysis, we identified an 11-Mb region on the X chromosome (Xp11.3-p11.22) that is linked with MN in these families. This region can potentially be narrowed to a 2-Mb (Xp11.22) locus at which all marker alleles are identical-by-state if identical-by-decent inheritance from a common ancestor is inferred. These results suggest that familial MN represents a different genetic etiology than the more commonly associated sporadic PLA2R-positive MN and that perhaps the inheritance is derived from an X-linked susceptibility locus mapped to this newly identified X-linked region.

X-Linked Familial MN

There are several reasons why the linked region identified on the X chromosome is convincing to explain the pattern and predominance of primary MN in our 3 families. First, 7 of 8 of the affected individuals presented with nephrotic syndrome in childhood. We know that MN in childhood is rare, accounting for only 1.5% to 7% of children undergoing biopsy for NS.[20–22] However, because many children with steroid-sensitive NS do not undergo biopsy, the true prevalence remains unclear. The incidence of childhood MN is estimated at less than 1 per 1,000,000-child population per year.[23] If we consider this estimated incidence and calculate the likelihood of these affected family members presenting with MN by chance (given the absence of known genetic risk alleles in the families), we find that likelihood would be less than 10^{-18} in families 1 and 2, and less

Table 1. Unique genes within 2 megabases of shared haplotype (Xp11.22)

AC239367.3, LINC01284, AC233976.1, AC233976.2, NUDT10, AL158055.1, EZHIP, NUDT11, LINC01496, CENPVL3, CENPVL2
CENPVL1, GSPT2, MAGED1, AC241520.1, RNU6-504P, AL929410.1, IPO7P1, AL929410.2, TPMTP3, AC239585.1, AC239585.2, MAGED4B, SNORA11E, MAGED4, SNORA11D, AC231759.2, AC245177.1, AC231759.1, MIR8088, XAGE2, AC231532.1, AC231532.2, BX510359.1, BX510359.5, BX510359.8, BX510359.7, RBM22P6, XAGE1A, BX510359.6, BX510359.2, BX510359.4
BX510359.3, SSXP4, SSXP1, AL450023.2, SSX8P, SSX7
AL450023.1, RNA5SP504, SSXP5, AL450023.3, SSX2, AC244505.2, AC244505.3, AC244505.5, SSX2B, AC244505.7
AC244505.4, SPANXN5, AC244505.6, XAGE5, AC244505.1
EIF4A2P4, XAGE3, AC244505.1, EIF4A2P4, FAM156B
AC234031.1, FAM156A

than 10^{-12} in family 3. Therefore, it is highly likely that these families share a common basis for disease.

Furthermore, MN is known to have a 2:1 male predominance. Whether the biological factors that contribute to this are related to the X-linked mechanism associated with the disease in the 3 families presented here is unknown. Importantly, previous genome-wide association studies analyses in MN did not include the X chromosome due to its unique statistical challenges, so they would not have detected any common genetic variants located there that contribute to disease risk[7–10,24]

Shared Haplotypes

Haplotype reconstruction further identified a 2-Mb identical shared haplotype in all individuals who carried the risk allele. This suggests that all 3 families share a distant common ancestor, representing a founder effect. Extended family histories were not available. Comparison of haplotypes between the individuals who were (n = 9) and were not (n = 17) carrying the 11-Mb risk allele revealed that no non-carrying individuals harbored this 2-Mb region, suggesting that this result is highly unlikely to occur by chance. Furthermore, we looked at the 1000 Genomes Project phase 3 data set and found that this haplotype was not present in any of these control individuals (n =2504, of which 670 are of European ethnicity).[25] This result demonstrates that the haplotype is not a common haplotype in the population and implies identical by descent inheritance in the 3 families of the 2-Mb region.

Genes of Interest in Linked Regions on X Chromosome

The 11-Mb region of Xp11.3-p11.22 contains 167 genes that code for proteins with a wide range of physiologic functions. Focusing on the 2-Mb common haplotype (Xp11.22), however, further narrows this list to 70 potential genes. To hypothesize about which genes were implicated in disease, we dissected the types of genes/proteins represented within this region. We used the UCSC Human Gene Sorter, a data-mining tool, to narrow our list of regional genes.[26] Idiopathic MN is characterized by IgG4 deposition in the glomerular basement membrane, and IgG antibodies are synthesized exclusively by B cells.[27] Therefore, we restricted our UCSC search to genes/proteins implicated in immune function, including expression in B and T lymphocytes.

Using the UCSC Human Gene Sorter to visualize gene expression *via* Genotype-Tissue Expression (GTEx),[28] we identified *GSPT2*, the G1-to-S phase transition 2 protein that mediates translation termination of a large protein product,[29] and *MAGED1*, the melanoma antigen (MAGE) family member D1 gene that regulates transcription factor complex formation,[30] to both have above-average expression in lymphocytes and whole blood, respectively. Previous reports have associated an Xp11.22 deletion encompassing these genes to be associated with intellectual disability and developmental delay; however, the consequences of loss of function in either of these genes has not been elucidated in humans or mouse models to date.[31] Perhaps one of these genes has a role in mediating familial MN. For completion, we also searched in the Human Kidney Cell Atlas for gene expression in podocytes, but we found none.[32]

Lastly, within the UCSC Genome Browser, we used the Gene Ontology tool to identify whether any of the 70 genes were cell-membrane proteins that could possibly be implicated as presenting antigens to the immune system. *FAM156A* was the only gene labeled as such, but showed very low tissue specificity in the Human Protein Atlas.[33] Future work through whole-genome sequencing and including additional cases is needed to further unpack these hypotheses, in order to refine the locus and thereby limit the coding or noncoding shared rare variants among affected patients.

Disease Response to Immunosuppression in Families 1 and 3

We have demonstrated that the genetic region Xp11.3-p11.22 is linked to the development of familial MN and propose that this underlying genetic locus causes disease susceptibility rather than monogenic pathogenicity. This is suggested for 3 reasons: (i) affected family members in families 1 and 3 demonstrated varying response to immunosuppression; (ii) there were varying ages of presentation across all affected individuals (ages 1 to >18 years); and (iii) the absence of a history of further affected family members, which would be expected, if this locus had 100% penetrance. These clinical aspects raise the hypothesis that instead of having a causal genetic variant leading to disease, perhaps instead, these families have a genetically conferred susceptibility that predisposes them to disease but requires other triggers for them to fully develop MN.

Previous susceptibility genes have been described. Perhaps the best-known example is *APOL1*, in which specific genetic variations are found only in individuals of African descent that lead to increased risk of developing multiple types of kidney disease.[34] More importantly, polymorphisms in the *TNFA* gene have been associated with susceptibility to idiopathic MN in adults.[35] Indeed, further uncovering of the genes and

their function within this X-linked region will aid in this search.

CONCLUSIONS

In summary, our study shows significant linkage of familial MN to chromosome Xp11.3-11.22. MN family members have a significantly lower GRS than individuals with more typical MN associated with anti-PLA2R antibodies, suggesting that genetic risk in familial MN is distinct and that it encompasses an X-linked susceptibility locus mapped to this X-linked region.

DISCLOSURES

All the authors declared no competing interests.

ACKNOWLEDGEMENTS

RK, DB, and MLD were supported by St. Peter's Trust for Kidney, Bladder, & Prostate Research. MLD was supported by the Kidney Research Scientist Core Education and National Training Program (KRESCENT) Post-Doctoral Fellowship from the Kidney Foundation of Canada.

SUPPLEMENTARY MATERIAL

Supplementary File (PDF)

Table S1. Histopathology features in familial membranous nephropathy.

Figure S1. Haplotype reconstruction of 11-Mb-linked region (rs3027452-rs2360739) on chromosome X for families 1 to 3.

REFERENCES

1. Couser WG, Cattran DC. Chapter 20: Membranous Nephropathy. In: Floege J, Johnson RJ, Feehall J, eds. *Comprehensive Clinical Nephrology*. Fourth Ed. The Netherlands: Elsevier, Amsterdam; 2010:248–259.
2. Ronco P, Debiec H, Gulati S. Chapter 20: Membranous Nephropathy. In: Geary DF, Schaefer F, eds. *Pediatric Kidney Disease*. Second Edition. Berlin, Heidelberg: Springer-Verlag; 2016:529–546.
3. Couser WG. Primary membranous nephropathy. *Clin J Am Soc Nephrol*. 2017;12:983–997.
4. Bockenhauer D, Debiec H, Sebire N, et al. Familial membranous nephropathy: an X-linked genetic susceptibility? *Nephron Clin Pract*. 2008;108:c10–c15.
5. Chen Y, Schieppati A, Chen X, et al. Immunosuppressive treatment for idiopathic membranous nephropathy in adults with nephrotic syndrome. *Cochrane Database Syst Rev*. 2014: CD004293.
6. De Vriese AS, Glassock RJ, Nath KA, et al. A proposal for a serology-based approach to membranous nephropathy. *J Am Soc Nephrol*. 2017;28:421–430.
7. Stanescu HC, Arcos-Burgos M, Medlar A, et al. Risk HLA-DQA1 and PLA(2)R1 alleles in idiopathic membranous nephropathy. *N Engl J Med*. 2011;364:616–626.
8. Cui Z, Xie L-J, Chen F-J, et al. MHC class II risk alleles and amino acid residues in idiopathic membranous nephropathy. *J Am Soc Nephrol*. 2017;28:1651–1664.
9. Le W-B, Shi J-S, Zhang T, et al. HLA-DRB1*15:01 and HLA-DRB3*02:02 in PLA2R-related membranous nephropathy. *J Am Soc Nephrol*. 2017;28:1642–1650.
10. Xie J, Liu L, Mladkova N, et al. The genetic architecture of membranous nephropathy and its potential to improve non-invasive diagnosis. *Nature Commun*. 2020;11:1–18.
11. Kanigicherla D, Gummadova J, McKenzie EA, et al. Anti-PLA2R antibodies measured by ELISA predict long-term outcome in a prevalent population of patients with idiopathic membranous nephropathy. *Kidney Int*. 2013;83:940–948.
12. Bockenhauer D, Feather S, Stanescu HC, et al. Epilepsy, ataxia, sensorineural deafness, tubulopathy, and KCNJ10 mutations. *N Engl J Med*. 2009;360:1960–1970.
13. Abecasis GR, Cherny SS, Cookson WO, Cardon LR. Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet*. 2002;30:97–101.
14. Gudbjartsson DF, Jonasson K, Frigge ML, Kong A. Allegro, a new computer program for multipoint linkage analysis. *Nat Genet*. 2000;25:12–13.
15. Rüschendorf F, Nürnberg P. ALOHOMORA: a tool for linkage analysis using 10K SNP array data. *Bioinformatics*. 2005;21: 2123–2125.
16. Tekman M, Medlar A, Mozere M, et al. HaploForge: a comprehensive pedigree drawing and haplotype visualization web application. *Bioinformatics*. 2017;33:3871–3877.
17. Nyholt DR. All LODs are not created equal. *Am J Hum Genet*. 2000;67:282–288.
18. Patel KA, Oram RA, Flanagan SE, et al. Type 1 Diabetes Genetic Risk Score: a novel tool to discriminate monogenic and type 1 diabetes. *Diabetes*. 2016;65:2094–2099.
19. Lee CM, Barber GP, Casper J, et al. UCSC Genome Browser enters 20th year. *Nucleic Acids Res*. 2020;48:D756–D761.
20. Churg J, Habib R, White RR. Pathology of the nephrotic syndrome in children: a report for the International Study of Kidney Disease in Children. *Lancet*. 1970;295:1299–1302.
21. Valentini RP, Mattoo TK, Kapur G, Imam A. Membranous glomerulonephritis: treatment response and outcome in children. *Pediatr Nephrol*. 2009;24:301–308.
22. Gulati S, Sengupta D, Sharma RK, et al. Steroid resistant nephrotic syndrome: role of histopathology. *Indian Pediatr*. 2006;43:55–60.
23. Filler G, Young E, Geier P, et al. Is there really an increase in non-minimal change nephrotic syndrome in children? *Am J Kidney Dis*. 2003;42:1107–1113.
24. Accounting for sex in the genome. *Nat Med*. 2017;23:1243.
25. Auton A, Abecasis GR, Altshuler DM, et al. A global reference for human genetic variation. *Nature*. 2015;526:68–74.
26. Kent WJ, Hsu F, Karolchik D, et al. Exploring relationships and mining data with the UCSC Gene Sorter. *Genome Res*. 2005;15:737–741.
27. Kuroki A, Shibata T, Honda H, et al. Glomerular and serum IgG subclasses in diffuse proliferative lupus nephritis, membranous lupus nephritis, and idiopathic membranous nephropathy. *Intern Med*. 2002;41:936–942.

28. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat Genet*. 2013;45:580–585.

29. Chauvin C, Salhi S, Le Goff C, et al. Involvement of human release factors eRF3a and eRF3b in translation termination and regulation of the termination complex formation. *Mol Cell Biol*. 2005;25:5801–5811.

30. Sullivan AE, Peet DJ, Whitelaw ML. MAGED1 is a novel regulator of a select subset of bHLH PAS transcription factors. *FEBS J*. 2016;283:3488–3502.

31. Grau C, Starkovich M, Azamian MS, et al. Xp11.22 deletions encompassing CENPVL1, CENPVL2, MAGED1 and GSPT2 as a cause of syndromic X-linked intellectual disability. *PLoS One*. 2017;12:e0175962.

32. Liao J, Yu Z, Chen Y, et al. Single-cell RNA sequencing of human kidney. *Sci Data*. 2020;7:4.

33. Uhlén M, Fagerberg L, Hallström BM, et al. Proteomics. Tissue-based map of the human proteome. *Science*. 2015;347:1260419.

34. Friedman DJ, Pollak MR. *APOL1* and kidney disease: from genetics to biology. *Annu Rev Physiol*. 2020;82:323–342.

35. Thibaudin D, Thibaudin L, Berthoux P, et al. TNFA2 and d2 alleles of the tumor necrosis factor alpha gene polymorphism are associated with onset/occurrence of idiopathic membranous nephropathy. *Kidney Int*. 2007;71:431–437.

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-020-15383-w OPEN

The genetic architecture of membranous nephropathy and its potential to improve non-invasive diagnosis

Jingyuan Xie et al.#

Membranous Nephropathy (MN) is a rare autoimmune cause of kidney failure. Here we report a genome-wide association study (GWAS) for primary MN in 3,782 cases and 9,038 controls of East Asian and European ancestries. We discover two previously unreported loci, *NFKB1* (rs230540, OR = 1.25, $P = 3.4 \times 10^{-12}$) and *IRF4* (rs9405192, OR = 1.29, $P = 1.4 \times 10^{-14}$), fine-map the *PLA2R1* locus (rs17831251, OR = 2.25, $P = 4.7 \times 10^{-103}$) and report ancestry-specific effects of three classical HLA alleles: *DRB1*1501* in East Asians (OR = 3.81, $P = 2.0 \times 10^{-49}$), *DQA1*0501* in Europeans (OR = 2.88, $P = 5.7 \times 10^{-93}$), and *DRB1*0301* in both ethnicities (OR = 3.50, $P = 9.2 \times 10^{-23}$ and OR = 3.39, $P = 5.2 \times 10^{-82}$, respectively). GWAS loci explain 32% of disease risk in East Asians and 25% in Europeans, and correctly re-classify 20–37% of the cases in validation cohorts that are antibody-negative by the serum anti-PLA2R ELISA diagnostic test. Our findings highlight an unusual genetic architecture of MN, with four loci and their interactions accounting for nearly one-third of the disease risk.

#A full list of authors and their affiliations appears at the end of the paper.

Membranous Nephropathy (MN) is a rare cause of kidney failure, manifesting as nephrotic syndrome with a peak incidence between 30 and 50 years of age[1]. The landmark discoveries of pathogenic antibodies against neutral endopeptidase in antenatal MN[2], and anti-phospholipase A2 receptor (PLA2R) antibodies in adult MN[3] have established MN as the disease of autoantibodies directed against podocyte antigens. Several studies have confirmed the presence of autoantibodies against PLA2R in ~60–70% of cases of primary MN[4], with another 3–5% potentially explained by antibodies against thrombospondin type 1 domain-containing 7A[5].

Previous genome-wide association study (GWAS) for MN conducted in 75 French, 146 Dutch and 335 British cases genotyped with low resolution arrays identified impressively strong associations of the *HLA* region and the *PLA2R1* locus encoding the dominant antigen in MN[6]. These findings suggest that genetic variation controls the immunogenicity and/or expression level of the PLA2R auto-antigen, as well as the production of anti-PLA2R autoantibodies in individuals with a permissive HLA haplotype. However, specific causal alleles underlying GWAS associations have not yet been mapped at high resolution. Moreover, prior GWAS was limited to Europeans, and the reported associations have not been examined comprehensively across different ethnicities. Lastly, because of small sample size, the prior study might have missed additional disease relevant loci.

Herein, we report a genetic study of primary MN involving 12,820 individuals (3782 biopsy-documented cases and 9038 ancestry-matched controls), across nine cohorts of East Asian and European ancestries. The composition of our cohorts reflects the demographics of the centres that have collected DNA samples for genetic studies of this rare disease over the past 15 years. By using high resolution arrays with genome-wide imputation and over 7-fold increase in sample size compared to the prior GWAS, we discover two previously unreported genome-wide significant risk loci for MN and perform high resolution mapping and ethnicity-specific analyses of the known loci.

We describe an unusual genetic architecture of MN, with four loci and their genetic interactions accounting for nearly one-third of the disease risk. Our study implicates dysregulation of *NFKB1* and *IRF4* genes in the disease pathogenesis, providing genetic support for potential targeting of the NF-κB and interferon signalling pathways in primary MN. We also refine ethnicity-specific effects at the *HLA* locus, defining *DRB1*1501* as a major risk allele in East Asians, *DQA1*0501* in Europeans, and *DRB1*0301* in both ethnicities. We describe a risk haplotype at the *PLA2R1* locus that has a regulatory function and exhibits strong genetic interactions with the *HLA-DRB1* risk alleles. Lastly, we calculate a genetic risk score (GRS) based on these findings which, when used in combination with a serum anti-PLA2R ELISA (a serologic test for MN currently in clinical use), shows superior performance in discriminating cases and controls than the ELISA or GRS alone. We validate the performance of this combined risk score (CRS) in external validation cohorts. Our results demonstrate that a combined serum-genetic test can potentially be used to establish a new diagnosis of primary MN, obviating the need for a high risk kidney biopsy procedure in the majority of cases.

Results

Study design. Our study involved nine case-control cohorts, including four East Asian cohorts of 4841 individuals (1632 primary MN cases and 3209 controls) and five European cohorts of 7979 individuals (2150 primary MN cases and 5829 controls). Eight cohorts were genotyped with high density SNP arrays, imputed using the latest whole genome sequence reference panels, and meta-analyzed genome-wide, and the top 46 loci selected based on $P < 5 \times 10^{-5}$ were tested by targeted genotyping in the ninth cohort (Supplementary Table 1). The summary of study cohorts, genotyping methods, and ancestry-specific imputation panels is provided in Table 1.

All cases used in this study were defined by a kidney biopsy diagnosis of idiopathic MN and any suspected secondary cases due to drugs, malignancy, infection, or autoimmune disease were excluded. With the exception of the German Chronic Kidney Disease (GCKD) cohort, all controls used for discovery involved healthy population controls and any individuals with a known diagnosis of kidney disease were excluded. The GCKD cohort was drawn entirely from a prospective observational study of patients with CKD and consisted of biopsy-defined cases and controls for whom CKD etiology was clearly assigned to a non-MN cause, as previously described[7].

All genome-wide significant loci ($P < 5 \times 10^{-8}$) were refined by cohort-stratified stepwise conditional analyses to define independently associated haplotypes. We also analyzed classical HLA alleles and all common amino acid polymorphisms at class I and class II genes imputed at high resolution. We performed detailed genomic annotations and explored epistatic effects for significant loci. Based on significant GWAS loci, we designed a GRS for MN

Table 1 Baseline characteristics of participants in the discovery and replication cohorts.

Cohort	Ancestry	No. of cases	No. of controls	Total	Genotyping platform	Imputation reference population panel
Asian Cohorts						
Chinese Discovery	East Asian	561	904	1465	Zhonghua-8 chip (Illumina)	1000G Phase 3 East Asians
Korean Discovery	East Asian	164	708	872	MEGA chip (Illumina)	1000G Phase 3 East Asians
Japanese Discovery	East Asian	81	358	439	MEGA chip (Illumina)	1000G Phase 3 East Asians
Chinese Replication	East Asian	826	1239	2065	KASP (targeted)	-
All Asian		1632	3209	4841		
European Cohorts						
European Discovery 1	European	611	1246	1857	MEGA chip (Illumina)	1000G Phase 3 Europeans
European Discovery 2	European	1045	1094	2139	MEGA chip (Illumina)	1000G Phase 3 Europeans
Turkish Discovery	European	254	336	590	MEGA chip (Illumina)	1000G Phase 3 Europeans
Sardinian Discovery	European	93	1498	1591	OmniExpress (Illumina)	1000G Phase 3 Europeans
GCKD Discovery	European	147	1655[a]	1802	Omni2.5Exome (Illumina)	HRC 1.1
All European		2150	5829	7979		
All Participants		3782	9038	12,820		

[a]GCKD participants with chronic kidney disease etiology assigned to a non-MN cause (see Methods).

and performed its validation in external cohorts, including the three previously published European GWAS cohorts[6] and the Nephrotic Syndrome Study Network (NEPTUNE) study[8].

The descriptions of all cohorts including ancestry analyses and details of statistical approaches are provided in Methods, Supplementary Methods, and Supplementary Figs. 1 and 2.

Genome-wide association. The results of combined genome-wide meta-analyses are summarized in Fig. 1 and Table 2, with more information provided in Supplementary Table 1 and Supplementary Figs. 3 and 4. We discovered two novel genome-wide significant loci: a locus on chromosome 4q24 encoding *NFKB1* (rs230540, OR = 1.25, Meta-analysis $P = 3.4 \times 10^{-12}$) and a locus

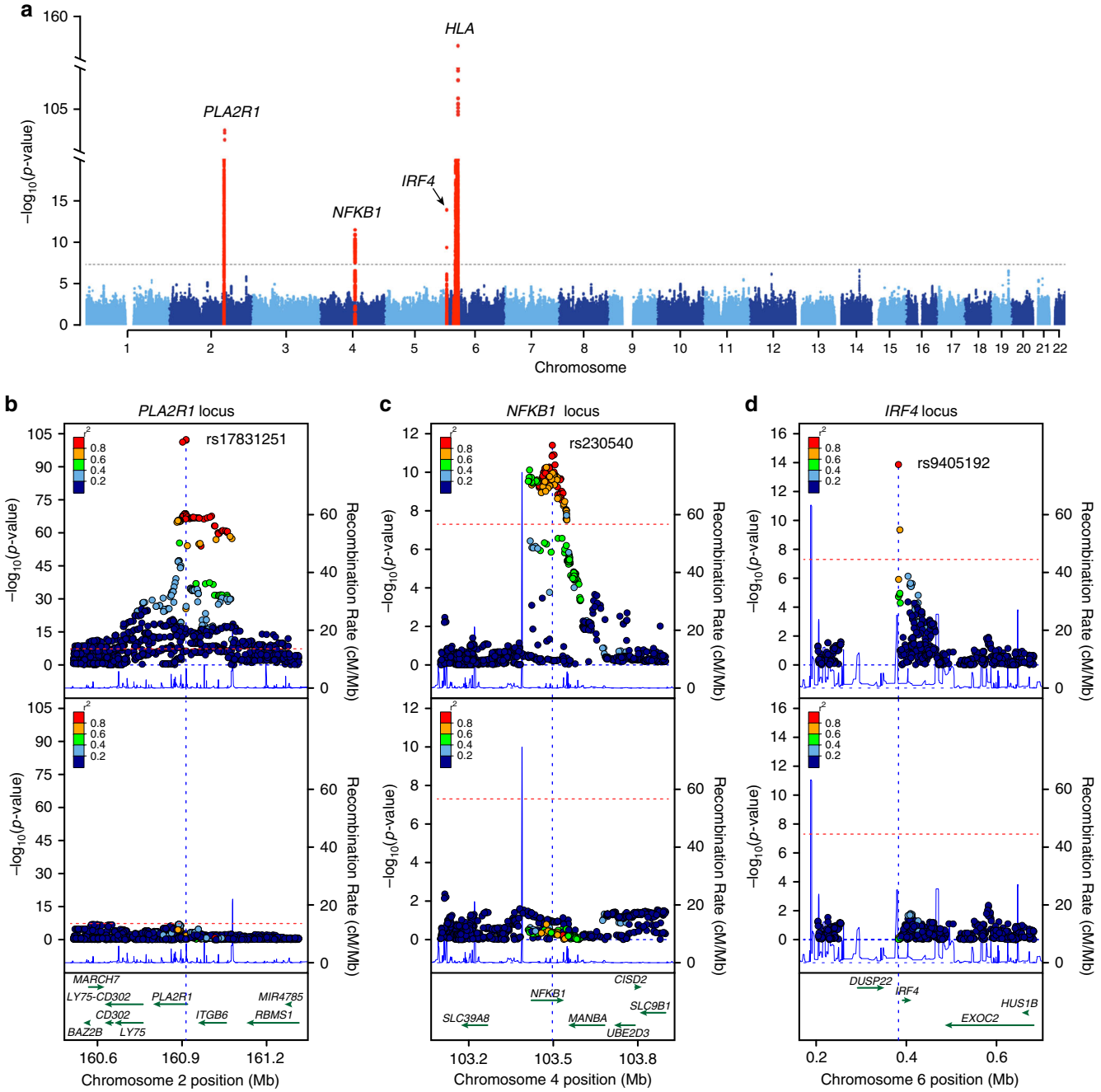


Fig. 1 Manhattan and regional plots for non-HLA loci for the combined meta-analysis of all MN cohorts. a The results of the combined meta-analysis across all cohorts; the dotted horizontal line indicates a genome-wide significance threshold ($\alpha = 5 \times 10^{-8}$); the *y*-axis is truncated twice to accommodate large peaks over *PLA2R1* and *HLA* loci; genome-wide-significant loci highlighted in red; **b** Regional plot for the *PLA2R1* locus; the upper panel shows unconditioned meta-results, the lower panel depicts meta-results after conditioning for the top SNP (rs17831251). **c** Regional plot for the *NFKB1* locus; the upper panel corresponds to unconditioned results; the lower panel shows meta-results after controlling for rs230540. **d** Regional plot for the *IRF4* locus; the upper panel corresponds to unconditioned results; the lower panel shows meta-results after controlling for rs9405192. The *x*-axis denotes genomic location (hg19 coordinates), left *y*-axis represents -log *P* values for association statistics, right *y*-axis represents average recombination rates based on HapMap-III reference populations combined (blue line). In the conditional analyses, we conditioned on the top SNP in each individual cohort, then meta-analyzed conditioned summary statistics as described in the Methods.

Table 2 Effect estimates for top GWAS SNPs by ethnicity and combined across all cohorts.

Locus	SNP	Risk allele	E. Asian case freq.	E. Asian control freq.	E. Asian OR (95% CI)	E. Asian *P*-value	European case freq.	European control freq.	European OR (95% CI)	European *P*-value	Combined OR (95% CI)	Combined *P*-value
PLA2R1	rs17831251	*C*	0.85	0.70	2.81 (2.48–3.17)	3.5×10^{-61}	0.76	0.61	1.98 (1.81–2.17)	4.7×10^{-48}	2.25 (2.09–2.42)	4.7×10^{-103}
NFKB1	rs230540	*C*	0.43	0.35	1.24 (1.14–1.36)	1.8×10^{-6}	0.35	0.32	1.25 (1.14–1.36)	7.8×10^{-7}	1.25 (1.17–1.33)	3.4×10^{-12}
IRF4	rs9405192	*G*	0.51	0.42	1.40 (1.28–1.53)	8.8×10^{-14}	0.73	0.69	1.18 (1.07–1.29)	6.6×10^{-4}	1.29 (1.21–1.37)	1.4×10^{-14}
HLA	rs9271573	*A*	0.60	0.35	2.97 (2.69–3.28)	3.7×10^{-102}	0.62	0.44	2.06 (1.89–2.25)	1.8×10^{-60}	2.41 (2.26–2.57)	2.7×10^{-154}

on chromosome 6p25.3 encoding *IRF4* (rs9405192, OR = 1.29, Meta-analysis $P = 1.4 \times 10^{-14}$). We also confirmed strong and highly significant associations at the previously described loci, including chromosome 2q24.2 encoding *PLA2R1* (rs17831251, OR = 2.25, Meta-analysis $P = 4.7 \times 10^{-103}$) and 6p21.32 encoding *HLA-DQA1/DRB1* genes (rs9271573, OR = 2.41, Meta-analysis $P = 2.7 \times 10^{-154}$).

Conditional analyses of the three non-HLA loci revealed that each signal is explained by a single SNP in each cohort, suggesting a single shared risk haplotype per locus in East Asian and European populations (Fig. 1b–d). To further test if the causal variants at these loci are likely shared between Europeans and East Asians, we performed 99% credible set analyses using summary statistics for each ancestry-defined subgroup and compared them with credible sets derived from the trans-ethnic meta-analysis. We confirmed that the predicted causal variants derived from the trans-ethnic analysis were largely overlapping with ancestry-specific results (Supplementary Fig. 5). In contrast, stepwise conditional analyses of SNPs at the HLA region revealed a complex pattern of association, with at least three independently genome-wide significant SNPs explaining the signal across all cohorts (Supplementary Table 2).

Given the complexity of the association signal at the HLA locus and known differences in linkage disequilibrium (LD) patterns by ancestry, we performed additional analyses of this region separately in East Asians and Europeans. In the conditional analyses of the East Asian cohorts, only two independently associated SNPs explained the entire signal at this locus (rs9269027 and rs1974461). In Europeans, stepwise conditional analyses revealed three independently associated genome-wide significant SNPs (rs9271541, rs9265949, and rs2858309), suggesting a more complex pattern of association (Supplementary Table 2). In both ethnicities, the top signal centred on *HLA-DRB1* and *DQA1* genes (Fig. 2a, b).

Classical HLA alleles and amino acid polymorphisms. We next imputed classical HLA alleles at two- and four-digit resolution using ethnicity-specific reference panels (see Methods). The first two digits specify a group of HLA alleles known as super-types as defined by older typing methodologies. The third through fourth digits specify nonsynonymous substitutions. Moreover, we imputed individual amino acid polymorphisms at class I (*HLA-A*, *-B*, and *-C*) and class II (*HLA-DQB1*, *-DQA1* and *-DRB1*) genes.

In East Asian cohorts, stepwise conditioning on classical HLA alleles defined two independent risk alleles, *DRB1*1501* (OR = 3.81, Wald test $P = 2.0 \times 10^{-49}$) and *DRB1*0301* ($OR_{conditioned}$ = 3.88, Wald test $P = 4.5 \times 10^{-24}$, Fig. 2c, Supplementary Table 3). In the analysis of polymorphic amino acid sites, genetic variation at only two codons encoding residues at positions 13 and 71 in DRβ, explained the entire *HLA-DRB1* signal (Fig. 3a, Supplementary Table 4). Specifically, DRβ position 13 occupied by Arginine (OR = 3.68, 95% CI: 2.74–4.95) or Serine (OR = 2.76, 95% CI: 2.06–3.71), and position 71 occupied by Lysine (OR = 3.10, 95% CI: 2.49–3.86) or Alanine (OR = 2.96, 95% CI: 2.55–3.45) conveyed the greatest risk (Supplementary Table 5). Consistent with a prior study in Chinese patients[9], these amino acids define the classical risk alleles *DRB1*1501* and *DRB1*0301* (Supplementary Table 6), and their side chains map adjacent to each other within the antigen-binding pocket of the β-chain of DR (Fig. 3c).

The top Asian risk allele *DRB1*1501* had no significant risk effect in Europeans despite its frequency being comparable between populations (control freq. 10% vs. 8% in Europeans and East Asians, respectively). The most strongly associated European risk allele was *DQA1*0501* (OR = 2.88, Wald test $P = 5.7 \times 10^{-93}$, Fig. 2d). After conditioning the locus on *DQA1*0501*, *DRB1*0301* remained genome-wide significant ($OR_{conditioned}$ = 2.00, Wald test $P = 2.0 \times 10^{-19}$, Supplementary Table 7) suggesting that this risk allele is shared between Asian and European populations. We note that *DQA1*0501* allele is twice as common in Europeans compared to Asians (control freq. 30% vs. 14%). Moreover, *DQA1*0501* and *DRB1*0301* are in imperfect LD that is stronger in Europeans ($r^2 = 0.40$) compared to East Asians ($r^2 = 0.29$). Although a weak effect of *DQA1*0501* was apparent in our Asian cohorts (Fig. 2c, Supplementary Table 3), this allele became non-significant after conditioning on *DRB1*0301*. In contrast, *DQA1*0501* exhibited a genome-wide significant risk effect after conditioning on *DRB1*0301* in Europeans ($OR_{conditioned}$ = 2.40, Wald test $P = 1.8 \times 10^{-18}$).

Given that our HLA imputation reference panels were considerably smaller for East Asians compared to Europeans, we sought additional validation of the observed classical HLA associations that were Asian-specific. We therefore created another reference panel based on the MHC sequence data from Zhou et al.[10] including 10,689 control individuals of Han Chinese ancestry. Using SNP2HLA software, we then re-imputed classical HLA alleles for our East Asian cohorts. We used the same quality control filters (MAF > 0.01 and imputation R^2>0.8) and methods for association testing as described above. We observed no major differences in the association statistics for the two Asian risk alleles, *DRB1*1501* (OR = 3.49, P = 3.85e−40) and *DRB1*0301* (OR = 4.08, P = 6.3E−24), demonstrating that these effects do not represent artifacts of smaller imputation panels.

We next performed the analysis of HLA amino acid substitutions in Europeans. Consistent with the association analyses of classical alleles, five bi-allelic sites in DQA1 that correlate with the *DQA1*0501* allele were most strongly associated with the risk of MN (75Ser-107Ile-156Leu-161Glu-163Ser, Wald test $P = 5.7 \times 10^{-93}$, Fig. 3b, Supplementary Tables 8–10). Conditioning on this haplotype in Europeans uncovered a second independent signal in *HLA-DRB1*, position 74 (Supplementary Tables 8 and 9), with Arginine representing the key risk residue (OR 2.86, 95% CI: 2.54–3.23). This residue defines the European *DRB1*0301* risk haplotype (Supplementary Tables 9 and 10). Notably, positions 74 (Europeans) and 71 (East Asians) are separated by a single turn along the α-helix, and their side chains are spatially close to that of position 13, located on the beta-sheet floor with its side chain oriented into the peptide-binding groove (Fig. 3c).

To confirm ethnicity-specific HLA effects, we repeated stepwise conditioning in a joint stratified analysis of all cohorts using bi-allelic tests of HLA alleles with formal tests of heterogeneity (Supplementary Table 11, Fig. 4a). The top classical allele supported by all cohorts regardless of ethnicity was *DRB1*0301*

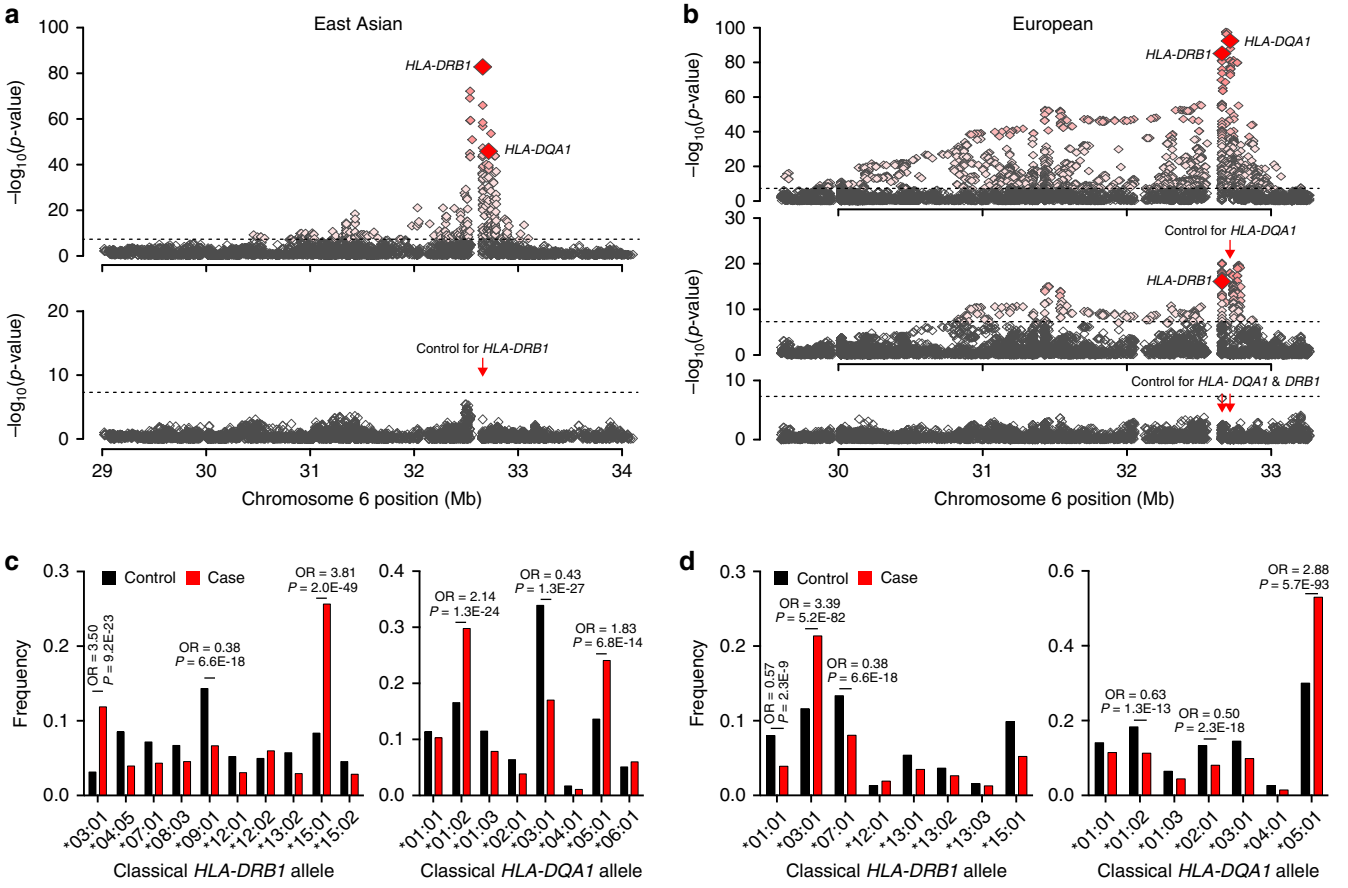


Fig. 2 Ethnicity-specific association analyses point to *HLA-DRB1* and *HLA-DQA1* as the top associated genes. a Regional plot for East Asian cohorts that includes association statistics for all imputed variants and classical HLA alleles; the strongest association signal was for *HLA-DRB1* gene (upper panel); the results for *HLA-DQA1* gene are highlighted for reference; after adjusting for all *DRB1* classical alleles (red arrow), there were no residual associations across the entire 3-Mb region (lower panel). The dotted horizontal line indicates the genome-wide significance threshold ($\alpha = 5 \times 10^{-8}$). **b** In Europeans, the strongest HLA gene association was for the *HLA-DQA1* gene (upper panel); after controlling for all classical *DQA1* alleles (red arrow), *HLA-DRB1* gene remained genome-wide significant (middle panel); after controlling for both *DRB1* and *DQA1*, there were no significant associations in the region (lower panel), suggesting that variation in both genes explains the entire signal. **c** East Asian frequency distributions of classical *DRB1* and *DQA1* alleles (four-digit resolution) for cases (red) and controls (black); unadjusted ORs and *P*-values provided for genome-wide significant alleles. **d** The European frequency distributions of classical *DRB1* and *DQA1* alleles (four-digit resolution) for cases (red) and controls (black); unadjusted ORs and *P*-values provided for genome-wide significant alleles.

(OR = 3.71, Wald test $P = 2.9 \times 10^{-127}$). After conditioning for *DRB1*0301*, the top classical allele was *DQA1*0501* ($OR_{conditioned} = 1.80$, Wald test $P = 1.1 \times 10^{-30}$), but this association was supported predominantly by Europeans. After controlling for both *DRB1*0301* and *DQA1*0501*, the top allele was *DRB1*1501* ($OR_{conditioned} = 1.94$, Wald test $P = 4.7 \times 10^{-29}$), but the risk effect was supported exclusively by East Asians (heterogeneity $I^2 = 97.5$, Cochrane's *Q*-test $P < 0.05$).

***PLA2R1* locus and its genetic interactions**. Consistent with prior GWAS, the most significant non-HLA locus resided on chromosome 2q24.2[6]. The top SNP was in the first intron of *PLA2R1*, which encodes the main podocyte autoantigen in primary MN. This signal was supported by both ethnicities, but the effect appeared stronger in East Asians (OR = 2.81, Meta-analysis $P = 3.5 \times 10^{-61}$) compared to Europeans (OR = 1.98, Meta-analysis $P = 4.7 \times 10^{-48}$, Table 2). After conditioning the association on the top SNP, rs17831251, there was no residual association at this locus, suggesting a common risk haplotype in both ethnicities (Fig. 1b).

We next refined the previously reported genetic interactions between the *PLA2R1* locus and HLA risk haplotypes. The *PLA2R1* risk genotype exhibited significant multiplicative interaction with both Asian and European HLA risk haplotypes (Fig. 4b, c), with the risk homozygosity at both loci associated with 89-fold increased odds of disease risk in East Asians [OR = 88.8 for double risk homozygotes (*N* cases/controls = 103/10) vs. double protective homozygotes (*N* cases/controls = 15/152), 95% CI: 38.0–207.3, Interaction test $P = 7.8 \times 10^{-3}$] and 14-fold in Europeans [OR = 14.1 for double risk homozygotes (*N* cases/controls = 291/89) vs. double protective homozygotes (*N* cases/controls = 52/237), 95% CI: 10.0–22.1, Interaction test $P = 6.4 \times 10^{-5}$]. Because the effect modification was weaker in Europeans, we next repeated interaction testing separating individual HLA risk haplotypes (Fig. 4d, e). Notably, the interaction in Europeans was driven predominantly by the *DRB1*0301-DQA1*0501* haplotype [OR = 28.7 for double risk homozygotes (*N* cases/controls = 115/13) vs. double protective homozygotes (N cases/controls = 75/339), 95% CI: 15.1–54.4, Interaction test $P = 2.2 \times 10^{-3}$]. After removing its effect, *DQA1*0501* had no residual interaction with *PLA2R1*

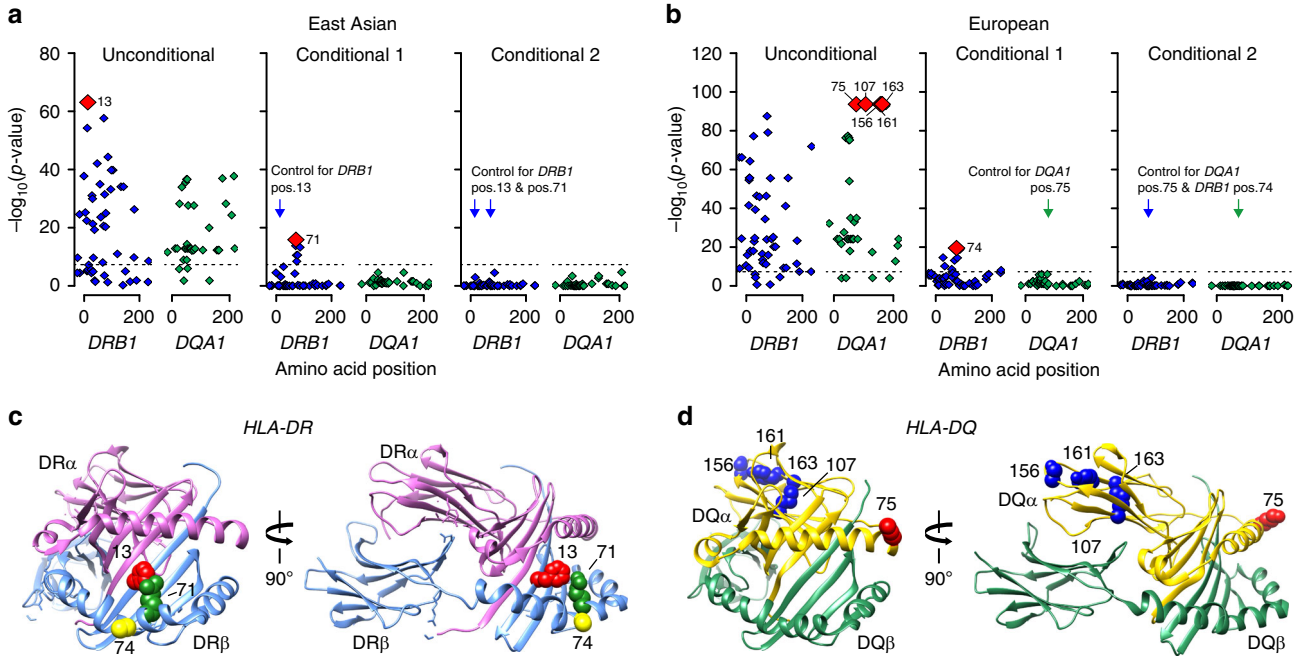


Fig. 3 Ethnicity-specific association analyses of DRβ1 and DQα1 amino acid sequence. a East Asian analysis of polymorphic amino acid positions within DRβ1 (blue) and DQα1 (green) molecules using conditional haplotype tests; the horizontal dash line marks the genome-wide significance level. The most strongly associated polymorphic site was position 13 in *DRB1* (left panel); after controlling for this position, position 71 in *DRB1* remained genome-wide significant (middle panel); after adjusting for both positions, there were no residual associations (right panel). **b** European analysis of polymorphic amino acid positions within DRβ1 (blue) and DQα1 (green); the DQA1 haplotype defined by amino acid positions 75, 107, 156, 161 and 163 (left panel) provided the strongest signal; after conditioning on this haplotype, position 74 in *DRB1* remained genome-wide significant (middle panel); no additional independent positions were found upon further conditioning (right panel). **c** Protein structure of the DR molecule including α chain (pink ribbon) and β chain (blue ribbon), the side chains of amino acids at DRβ1 positions 13, 71, and 74 are located adjacent to each other in the P4 pocket of the peptide-binding groove. **d** Protein structure of the DQ molecule including α chain (yellow ribbon) and β chain (green ribbon); all five bi-allelic amino acid sites in DQα1 are in perfect LD with each other and define the top associated *DQA1*0501* allele in Europeans.

(Interaction test $P = 0.1$). Similarly, there was no significant interaction between *DQA1*0501* allele and *PLA2R1* locus in East Asians. These analyses suggest that the *PLA2R1* locus interactions are driven primarily by the *DRB1* alleles.

We annotated all SNPs in LD with rs17831251 for potential impact on the structure and/or transcriptional regulation of *PLA2R1*. We found two common missense variants in moderate LD, rs35771982 (p.H300D, $r^2 = 0.69$) and rs3749117 (p.M292V, $r^2 = 0.68$), but the effects of these variants were considerably weaker compared to rs17831251, suggesting that they are unlikely to represent causal variants (Supplementary Table 12). Our tissue-specific functional scoring method for non-coding variants based on the ENCODE and Roadmap Epigenetics data[11] prioritized another variant in intron 1, rs17241973 ($r^2 = 0.93$ with rs17831251) that intersects a putative enhancer element across multiple tissues (Supplementary Fig. 6). Both rs17831251 and rs17241973 exhibit strong cis-eQTL effects on *PLA2R1* expression wherein the MN risk alleles associate with lower mRNA expression of *PLA2R1* across multiple tissues in GTEx[12], but this effect appears reversed for the kidney tissue (Supplementary Fig. 7). To further confirm these kidney-specific effects, we used gene expression data from manually micro-dissected human kidney compartments of 166 NEPTUNE participants[13]. We detected suggestive glomerular eQTL effects that were weak, but direction-consistent with GTEx for rs17831251 (Wald test $P = 0.055$) and rs17241973 (Wald test $P = 0.024$), wherein MN risk allele were associated with increased glomerular *PLA2R1* mRNA levels (Supplementary Fig. 8).

Because kidney tissue compartments are not well represented in either ENCODE or Roadmap datasets, we next examined the genomic location of rs17831251 and rs17241973 in relationship to the recently published kidney compartment-specific chromatin landscape[14] (Supplementary Fig. 9); rs17241973 lies within intron 1 of *PLA2R1* in open chromatin that is active in both glomerular and tubular compartments and is contiguous with the gene promoter. In contrast, rs17831251 lies within a broad region of increased chromatin accessibility in glomeruli, and is only 2.1-kb away from a glomerulus-specific DHS that contains a high-confidence NFKB1-binding motif. In glomerulus-specific chromatin conformation (Hi-C) data, both rs17831251 and rs17241973 make regional and distal contacts with other glomerular DHS, emphasizing a composite cis-regulatory module for *PLA2R1* gene expression.

Novel loci encoding *NFKB1* and *IRF4*. The 4q24 locus contains the *NFKB1* gene, which encodes an active DNA binding subunit of the NF-κB transcriptional complex. The top SNP, rs230540 (OR = 1.25, Meta-analysis $P = 3.4 \times 10^{-12}$), is an intronic variant predicted to have a functional effect specific to immune cells (Supplementary Fig. 10). In agreement with our prediction, rs230540 has been associated with higher mRNA expression of *NFKB1* in whole blood ($P = 2.6 \times 10^{-11}$)[15] and in CD4$^+$ T cells ($P = 2.0 \times 10^{-9}$)[16]. Consistent with pro-inflammatory effects of NF-κB, the MN risk haplotype at this locus determined higher leukocyte counts[17] and increased risk of ulcerative colitis[18,19] and primary biliary cholangitis[20,21] (Supplementary Table 13, Fig. 5a). *NFKB1* is also expressed in human podocytes[22], as well as primary human glomerular and tubular epithelial cell cultures[14], and rs230540 intersects an active glomerular DHS with several

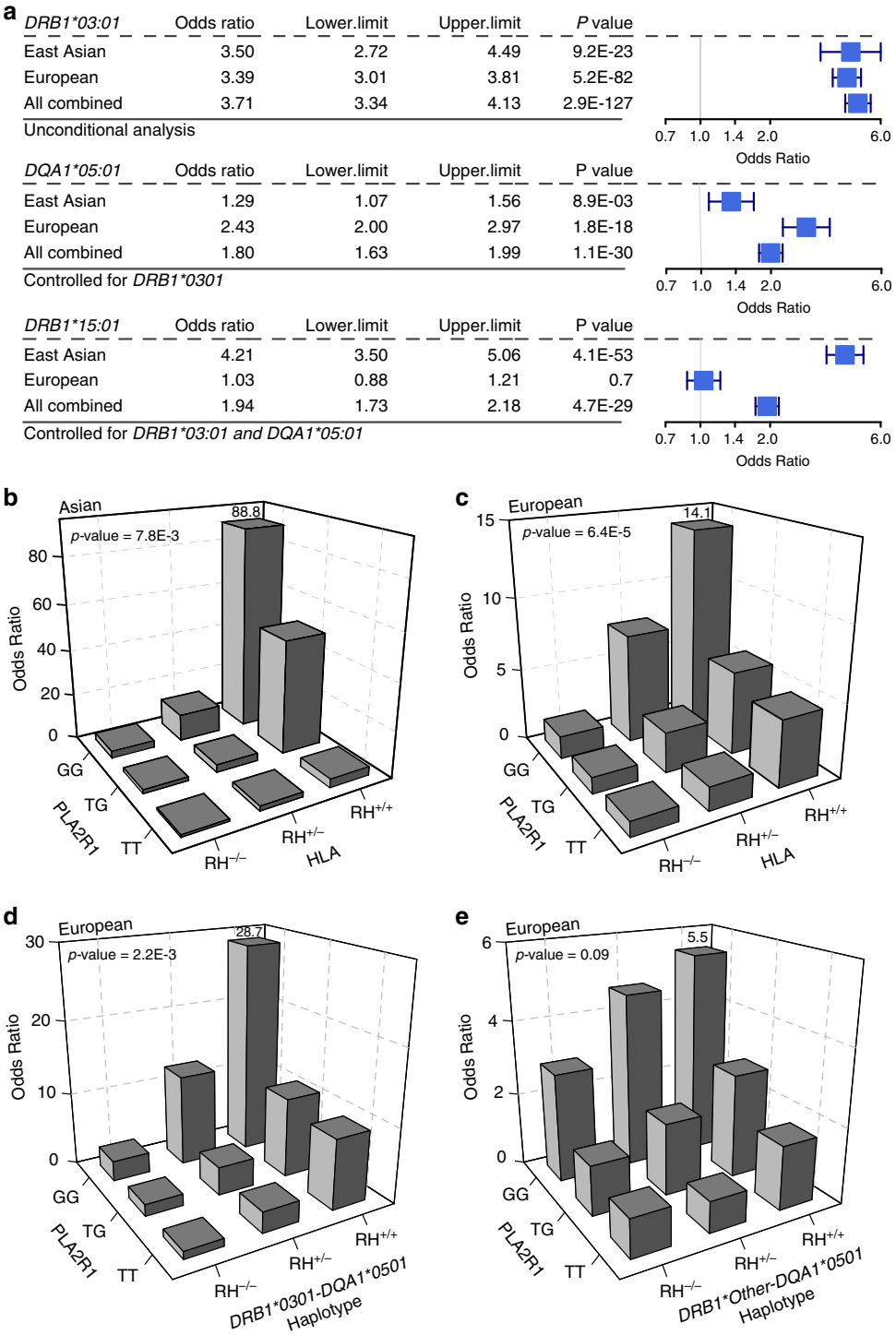


Fig. 4 Ethnicity-specific HLA allelic effects and genetic interactions. a Stepwise conditioning of HLA risk alleles by ethnicity; Top**:** Unconditioned effect estimates, 95% confidence intervals, and *P*-values for *DRB1*0301* demonstrate similar effects in East Asian and European cohorts. Middle: Effect estimates for *DQA1*0501* after controlling for *DRB1*0301* demonstrate significant heterogeneity, with stronger effects in the European cohorts (Cochrane's Q *P*-value < 0.05). Bottom: Effect estimates for *DRB1*1501* after controlling for both *DRB1*0301* and *DQA1*0501* demonstrate significant heterogeneity with risk effect in East Asians but no effect in Europeans (Cochrane's *Q P*-value < 0.05). **b** *PLA2R1* genotype interaction with East Asian HLA risk haplotypes, *DRB1*0301* or *DRB1*1501* (N cases/controls = 803/1,956, $P = 7.8 \times 10^{-3}$). **c** *PLA2R1* genotype interaction with European HLA risk haplotypes, *DRB1*0301* or *DQA1*0501* (*N* cases/controls = 1880/2627, multiplicative interaction test $P = 6.4 \times 10^{-5}$). **d** *PLA2R1* genotype interaction with *DRB1*0301-DQA1*0501* risk haplotype in Europeans (multiplicative interaction test $P = 2.2 \times 10^{-3}$). **e** No significant interaction between *PLA2R1* and *DQA1* risk haplotypes other than *DQA1*0501-DRB1*0301* in Europeans (multiplicative interaction test $P = 0.09$). RH risk haplotype.

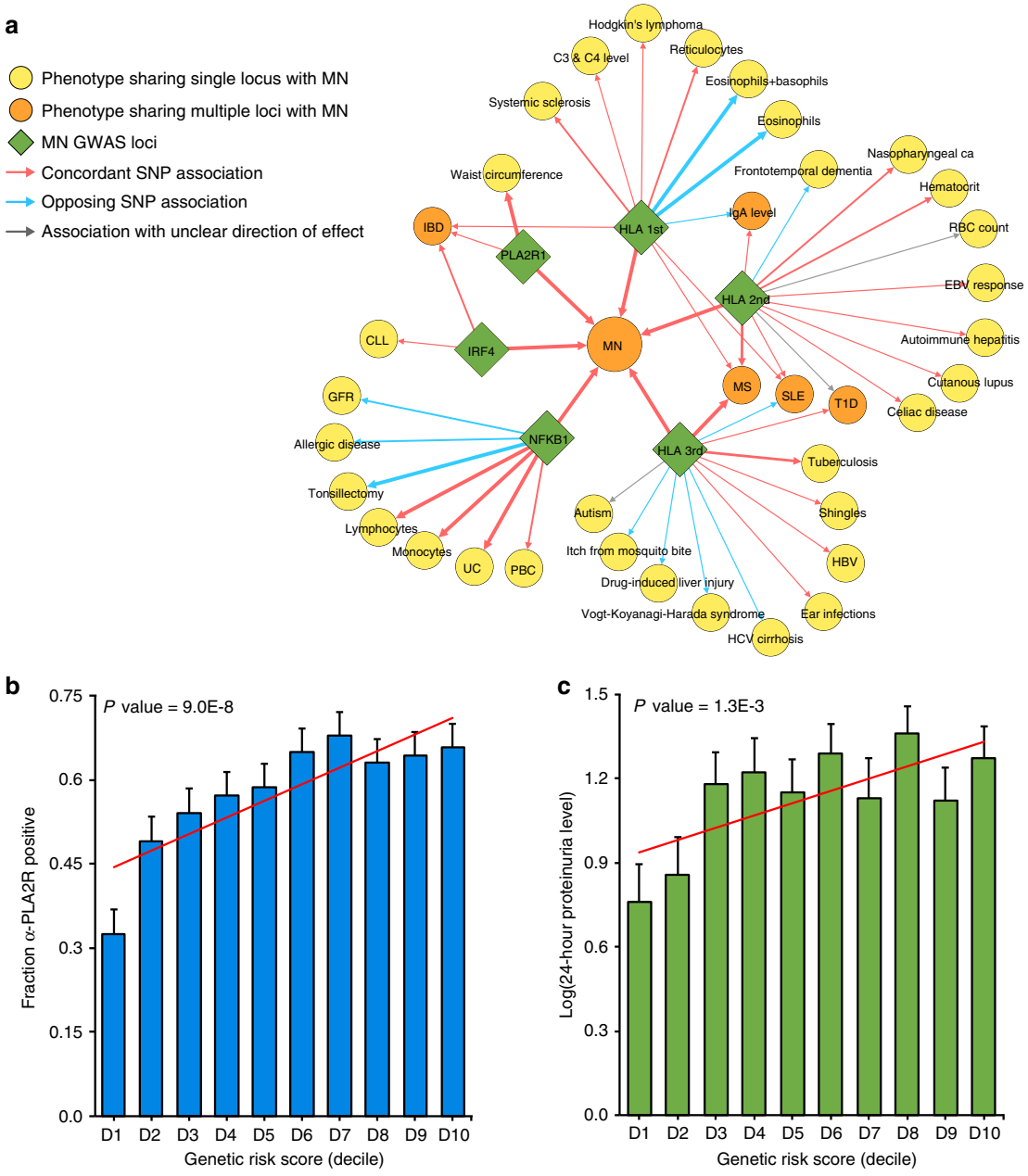


Fig. 5 Pleiotropic effects of the MN loci and their clinical correlations. a The pleiotropy map was constructed based on overlapping genome-wide significant loci reported in the GWAS Catalogue: traits sharing a single locus with MN are indicated in yellow; traits sharing multiple loci are indicated in orange; arrows represent allelic associations that are identical to, or in tight LD ($r^2 > 0.8$) with the MN risk alleles; arrow thickness is proportional to r^2 between alleles; concordant effects are indicated in red and opposed effects in blue. IBD: inflammatory bowel disease, includes ulcerative colitis (UC) and Crohn's disease (CD); PBC primary biliary sclerosis, GFR glomerular filtration rate, CLL chronic lymphocytic leukemia, HBV hepatitis B virus; **b** Significant positive correlation of the GRS with anti-PLA2R antibody seropositivity ($N = 1114$ cases, Wald test $P = 9.0 \times 10^{-8}$) and **c** log-transformed 24-h proteinuria at diagnosis ($N = 1329$ cases, Slope test $P = 1.3 \times 10^{-3}$). The *x*-axis depicts deciles of GRS; error bars correspond to standard errors; the *P*-values are adjusted for age, sex, and ethnicity.

glomerular Hi-C contact sites[14] (Supplementary Fig. 11). Notably, the MN risk allele has previously been associated with lower estimated glomerular filtration rate in GWAS of renal function[23,24], thus this locus may be more broadly associated with the risk of kidney disease.

The top SNP on chromosome 6p25.3, rs9405192 (OR = 1.29, Meta-analysis $P = 1.4 \times 10^{-14}$), resides upstream of *IRF4* gene, which belongs to the family of transcription factors regulating interferon-inducible genes. IRF4 is lymphocyte specific and negatively regulates Toll-like-receptor signalling that is central to the activation of innate immune system; this gene is known to be under the transcriptional control of the NF-kB complex[25–27]. Unlike *PLA2R1* and *NFKB1*, *IRF4* does not appear to be expressed in human kidney cells by single nuclei RNA-seq[28] (Supplementary Fig. 12). We did not find functional or coding SNPs in LD with rs9405192, nor did we observe any cis-eQTL effects for this variant, thus the precise mechanism underlying this association remains unknown. However, the analysis of binding sites for individual components of the NF-κB complex in lymphocytes[29] suggested binding of the complex in close proximity of rs9405192 (Supplementary Fig. 13). The risk allele at this locus is also in strong LD with variants previously

associated with increased risk of inflammatory bowel disease[19], and in weaker LD with several risk variants for chronic lymphocytic leukemia (Supplementary Table 13, Fig. 5a), suggesting the pattern of pleiotropy that is similar to the *NFKB1* locus. Nevertheless, we detected no statistically significant genetic interactions of *IRF4* and *NFKB1* loci.

In addition, we systematically annotated all other suggestive non-HLA loci defined by $P < 5.0 \times 10^{-5}$ and these results are summarized in Supplementary Table 14. To enhance potential genetic discovery of novel podocyte antigens, we also repeated genome scans after conditioning for the *PLA2R1* locus, but detected no additional suggestive loci.

SNP-based heritability and risk explained by GWAS. Using our genotype data and genome-based restricted maximum likelihood method (GREML)[30], we estimated the overall SNP-based heritability of MN at 0.43 (SE = 0.039) in East Asians and 0.36 (SE = 0.0046) in Europeans. Remarkably, all genome-wide significant risk alleles exhibited unusually large effect sizes for GWAS. In order to quantify the fraction of disease variance cumulatively explained by genome-wide significant SNPs and their interactions, we performed ethnicity-specific GRS analyses (see Methods). Each GRS was expressed as a weighted sum of risk alleles with weights defined by their mutually adjusted effect estimates and included the 3 independent non-HLA SNPs (rs6707458, rs230540, rs9405192) as well as ethnicity-specific HLA risk alleles and their interactions. This included rs9269027, rs1974461, and rs9269027*rs6707458 interaction term for East Asians, and rs9271541, rs9265949, rs2858309 and rs9271541*rs6707458 interaction term for Europeans (Supplementary Table 15). The GRS calculated using this method explained 32% disease risk in East Asians, 25% in Europeans, and 29% of overall disease risk across all cohorts combined. Remarkably, the magnitude of the GRS effect was comparable to rare, highly penetrant mutations causing Mendelian forms of kidney disease, with individuals in the top decile of GRS having 30 to 40-fold higher disease risk compared to the lowest decile (Fig. 6a, b).

Clinical correlations of the GRS. For a subset of patients with available clinical data, we performed genetic correlation analyses with selected clinical features reflective of disease severity. The GRS was positively correlated with PLA2R antibody seropositivity (Wald test $P = 9.0 \times 10^{-8}$), and in those with detectable antibodies, higher titers at the time of biopsy (Slope test $P = 1.2 \times 10^{-9}$, Fig. 5b). The GRS also predicted worse proteinuria at the time of biopsy, which represents the key marker of MN severity and prognosis (Slope test $P = 1.3 \times 10^{-3}$, Fig. 5c). Other clinical features, such as age at diagnosis, renal function, or serum albumin levels at the time of biopsy were not significantly correlated with the GRS after multivariate adjustment (Supplementary Table 16).

Potential diagnostic implications of the GRS. Although the diagnosis of MN is traditionally established by a kidney biopsy, the detection of circulating PLA2R antibodies by ELISA has recently emerged as a useful diagnostic modality[31]. In this study, we performed ELISA in sera obtained within 6 months of a diagnostic kidney biopsy in a total of 2331 individuals (1488 cases, 300 healthy controls, and 543 disease controls). In East Asians, we estimated that the standard ELISA cut-off of 20 U/mL provided 100% specificity and 60% sensitivity for the diagnosis of MN. In the analysis of Europeans, depending on the specific cohort, the same cut-off provided 99–100% specificity and 51–57% sensitivity (Supplementary Table 17). While the antibody level of 20 U/mL represents the manufacturer's recommended cut-off, levels 2–20 U/mL are frequently considered as borderline-negative, and levels <2 U/mL as negative[31]. In our cohorts, the cut-off 2 U/mL had inadequate diagnostic specificity (range 73–92%). These results confirm the key limitation of the PLA2R antibody ELISA, which has high specificity (99–100%) but low sensitivity (51–60%) at the standard recommended cut-off point; while lowering the cut-off increases sensitivity, it results in inadequate specificity. Consequently, the levels in the borderline-negative range (2–20 U/mL) are difficult to interpret clinically.

Given this limitation, we evaluated if the addition of genetic risk information can improve the performance of ELISA, especially in cases that fall in the borderline-negative range. First, we evaluated diagnostic properties of the GRS alone in our discovery cohorts. In East Asians, the genetic test had area under the receiver operating characteristics curve (AUROC) of 0.80 (95% CI: 0.78–0.82), while in Europeans the AUROC was 0.75 (95% CI: 0.74–0.77). Combining genetic and serologic tests in the form of a CRS provided superior case discrimination with AUROCs of 0.96 (95% CI: 0.95–0.98) in East Asians and 0.89 (95% CI: 0.87–0.91) in Europeans (Fig. 6, Supplementary Table 18).

We next tested the GRS performance in several external validation cohorts, including three independent GWAS cohorts of European ancestry as well as in the European-American NEPTUNE participants with incident nephrotic syndrome (Supplementary Table 18C-E). Overall, the effects and the diagnostic performance of the GRS were comparable between the European discovery and each validation cohort (Fig. 6c). When combined with the serum antibody titer, the CRS achieved AUROC of 0.96 (95% CI: 0.94–0.97) across all validation cohorts combined (Supplementary Table 18E, Fig. 6f).

In the subgroup analyses, we compared the diagnostic properties of GRS and CRS by the antibody status in all cohorts pooled by ancestry (Fig. 7). These analyses demonstrated that both GRS and CRS were predictive of case status even for antibody-negative MN. Importantly, the CRS continued to have excellent performance in classifying the borderline-negative cases (antibody level 2–20 U/mL range), with AUROCs of 0.98 (95% CI: 0.97–0.99) in East Asians and 0.95 (95% CI: 0.93–0.96) in Europeans. Notably, among all cases for which the ELISA test was either negative or inconclusive, adding genetic information in the form of CRS can establish the diagnosis in 20–37% cases with 99% specificity. The comparison of AUROCs between GRS, CRS, and serum anti-PLA2R Ab test by ancestry is provided in Supplementary Fig. 14, and the clinical implications of these findings are summarized in Supplementary Note 1 and Supplementary Table 19.

Lastly, we expanded our validation studies to non-European participants of the NEPTUNE study (Supplementary Tables 20 and 21). Although the European risk score performance was diminished in Hispanic Americans, the European GRS performed well in African Americans, and this is despite substantial differences in risk allele frequencies between Europeans and Africans (Supplementary Table 22). Similar to the European validation cohorts, the European GRS was superior compared to the trans-ethnic GRS when applied to the NEPTUNE minority populations, while the Asian GRS had relatively poor performance in both African-American and Hispanic/Latino cohorts (Supplementary Table 21, Supplementary Fig. 15).

Discussion

Our study provides important insights into an autoimmune disease and the genetic architecture of MN. First, we discover novel genome-wide significant risk loci for MN with large effects encoding two transcriptional master regulators of inflammation, NFKB1 and IRF4. The association at the *NFKB1* locus highlights

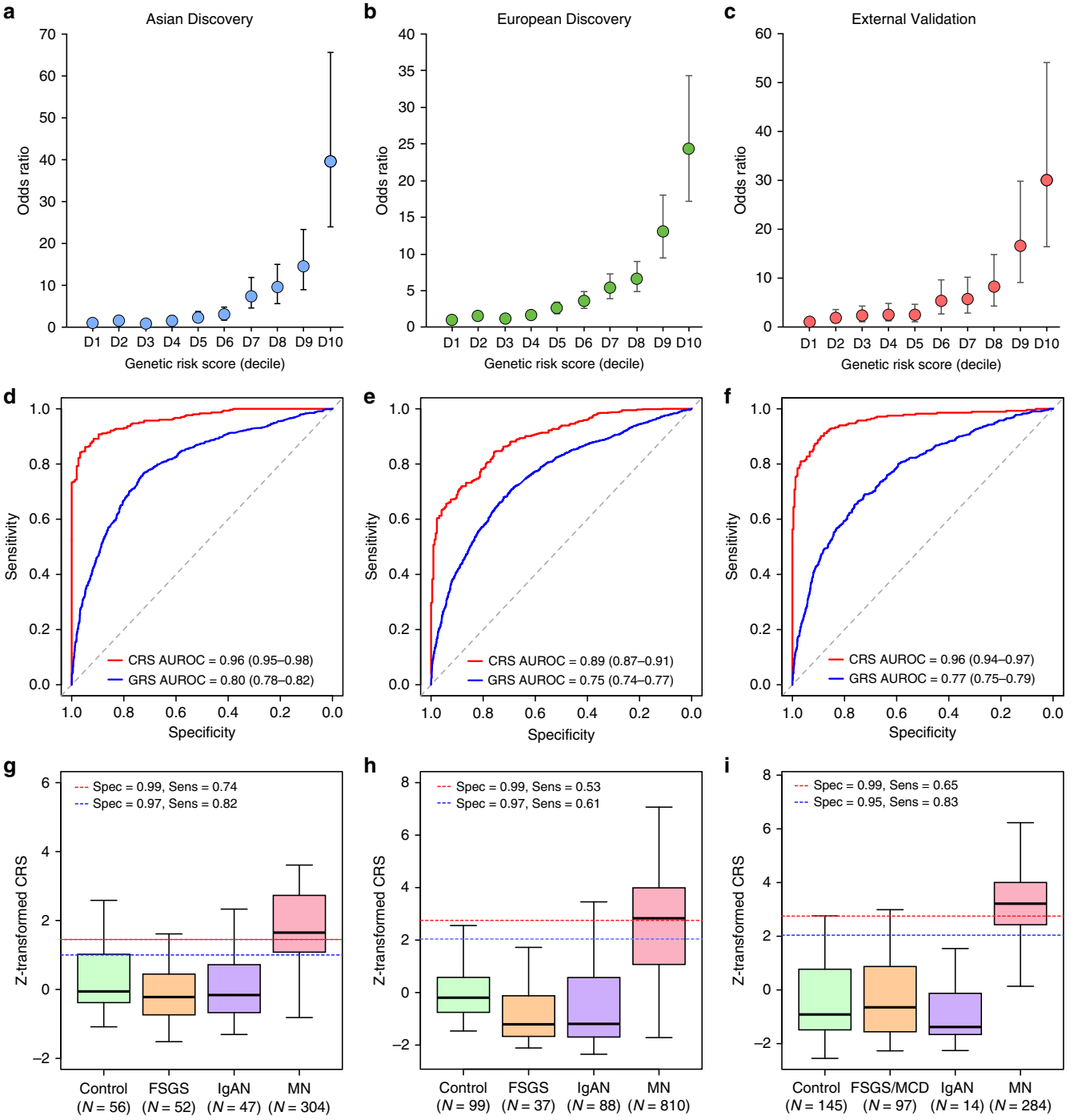
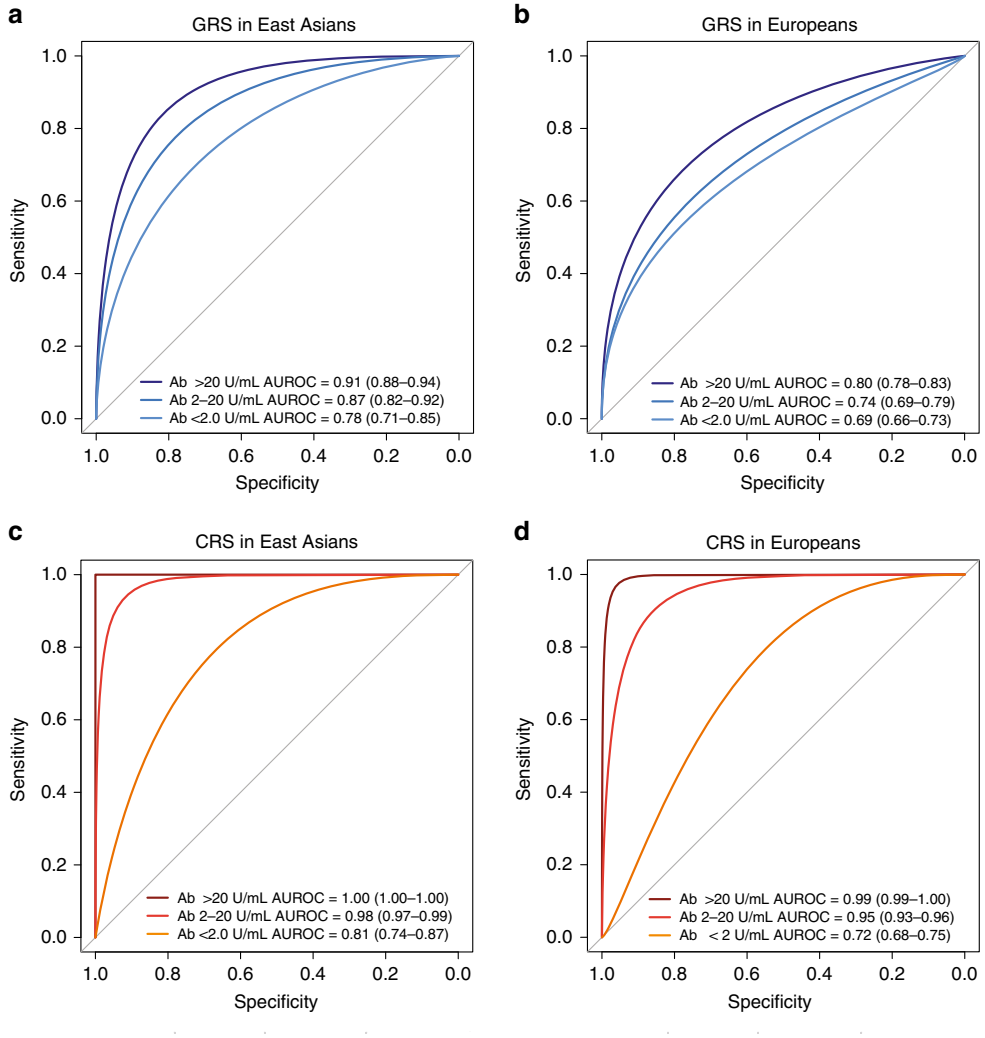


Fig. 6 Diagnostic performance of the genetic risk score (GRS) and the combined risk score (CRS). Genetic effects expressed as odds ratios (OR) and 95% confidence intervals in reference to the lowest decile of the GRS distribution for **a** East Asian discovery, **b** European discovery, and **c** European validation cohorts combined. GRS and CRS Receiver Operating Characteristics (ROC) curves for **d** East Asian discovery, **e** European discovery, and **f** European validation cohorts combined; AUROC Area under the ROC curve. Distributions of the CRS between healthy controls, diseased controls, and MN cases for **g** East Asian discovery, **h** European discovery, and **i** European validation cohorts combined. The box plots depict medians (horizontal lines), interquartile ranges (boxes), and minimum/maximum values (whiskers). The discovery CRS cut-offs of 1.00 and 1.45 in East Asians and 2.05 and 2.72 in Europeans have 97% and 99% specificity, respectively. The same European cut-offs were applied to the validation cohorts for comparisons of specificity and sensitivity.

the role of the canonical NF-κB pathway in primary MN. Upon activation by pro-inflammatory signals, NFKB1 undergoes proteasome processing to p50, an active DNA binding subunit of the NF-κB complex. Inappropriate activation of this pathway has previously been studied in progressive diabetic nephropathy[32], and in the context of inflammatory diseases, including IBD[33,34] and MN[35–37]. Importantly, the MN risk allele at this locus has a concordant effect on the risk of ulcerative colitis[18,19] and primary biliary cirrhosis[20,21]. It has also been associated with increased mRNA expression of *NFKB1* in cis, and reduced DNA



e East Asians

Subgroup	Cut-off	Specificity	Sensitivity
CRS All	1.45	99%	74%
CRS Ab-	1.45	99%	37%
GRS All	2.78	99%	15%

f Europeans

Subgroup	Cut-off	Specificity	Sensitivity
CRS All	2.50	99%	60%
CRS Ab-	2.50	99%	20%
GRS All	2.64	99%	13%

Fig. 7 Diagnostic properties of the genetic risk score (GRS) and combined risk score (CRS) stratified by anti-PLA2R antibody status. Comparisons of receiver operating characteristic (ROC) curves to discriminate antibody positive (PLA2R Ab > 20 U/mL), borderline negative (PLA2R Ab 2–20 U/mL), and negative (PLA2R Ab < 2 U/mL) cases of primary MN from all available healthy and diseased controls combined for **a** GRS in East Asian discovery cohorts, **b** GRS in European discovery and validation cohorts, **c** CRS in East Asian discovery cohorts, **d** CRS in European discovery and validation cohorts. Overall sensitivities for risk score cut-offs corresponding to 99% specificities in (**e**) East Asian discovery cohorts and (**f**) European discovery and validation cohorts; CRS All: all patients with serum Ab measurements by ELISA within 6 months of a diagnostic kidney biopsy; CRS Ab-: patient subgroup with PLA2R Ab < 20 U/mL, and GRS All: all patients with GWAS data available for GRS calculation. AUROC: area under the ROC curve (95% confidence interval).

methylation in trans across >400 CpGs that overlap with NF-κB-binding sites, suggesting enhanced baseline activity of NF-κB[38]. The NF-κB complex is known to up-regulate IRF4 expression with cross-regulatory feedback loops between *NFKB1* and *IRF4* described in several prior studies[25–27]. Taken together, NFKB1 and IRF4 loci participate in a common regulatory pathway in immune cells, and our genetic findings clearly establish a critical role of this pathway in the pathogenesis of MN.

Second, due to the bi-ethnic composition of our cohorts, we were able to refine ethnicity-specific effects at the *HLA* locus, defining *DRB1*1501* as a major risk allele in East Asians, *DQA1*0501* in Europeans, and *DRB1*0301* in both ethnicities. These findings suggest that different epitopes are likely presented to T cells to initiate the anti-PLA2R response in East Asians and Europeans. We also identified specific high-risk amino acid substitutions, at positions 13, 71, and 74, mapping to the P4 pocket of DRβ1. Although the same positions contribute to the risk of T1D[39] and rheumatoid arthritis[40], the effects of individual residues at each position are discordant, likely reflecting differences in target epitopes.

Third, we confirm that a single haplotype at the *PLA2R1* locus conveys the disease risk in both East Asians and Europeans, and exhibits genetic interactions with *HLA-DRB1* risk alleles. Our analysis supports a regulatory function of the *PLA2R1* risk

haplotype. The candidate causal variant resides in the first intron of *PLA2R1* and intersects a predicted enhancer element. While this variant is normally associated with suppressed *PLA2R1* transcription across multiple tissues, it appears to increase expression of *PLA2R1* in the kidney. This finding highlights the importance of studying target tissues and is consistent with the findings that among CKD loci that are transcriptionally active in renal tissue, 15.8% of effects are kidney-specific[41]. Notably, the top variants at the *PLA2R1* locus also intersect a putative NF-κB binding site in lymphocytes, although no similar data is presently available for podocytes. Further experimental work is thus needed to test if the glomerular-specific eQTL effect is under the transcriptional control of NF-κB. Moreover, larger glomerular compartment-specific datasets will be needed to confirm the observed eQTL effects.

Another observation is that all four genome-wide significant risk loci (*PLA2R1, IRF4, NFKB1*, and *HLA*) exhibit highly pleiotropic effects and all four lead SNPs have a concordant effect on the risk of inflammatory bowel disease (IBD). This observation suggests shared pathogenic mechanism between IBD and MN. Considering that MN is an orphan disease without a targeted treatment, there may now be opportunities for drug repositioning approaches from IBD, where several new anti-inflammatory agents are currently under development. Our study suggests that the NF-κB and interferon pathways may represent particularly attractive drug targets.

Remarkably, our GWAS loci are highly predictive of the disease status and jointly explain up to one third of disease risk, an exceptionally large fraction for common alleles. This may be partially explained by the fact that MN frequently occurs after the peak reproductive age, allowing the risk alleles to escape purifying selection. Moreover, even though the risk alleles are common, our interaction analysis demonstrates that specific high-risk genotype combinations are relatively rare in the general population, potentially explaining the low overall prevalence of MN[42]. The alternative hypothesis is that of balancing selection. NF-kB and IRF4 are both involved in immune defenses against common pathogens and some Phospholipase A2 ligands for PLA2R1 represent downstream NF-kB targets with antibacterial properties[43,44]. Therefore, the observed high frequencies of MN risk alleles could be explained by their protective effects against common infections.

Finally, a simple GRS based on our GWAS loci has excellent discriminant properties when combines with anti-PLA2R ELISA test. Importantly, the combined genetic-serum test has superior diagnostic properties compared to serologic test alone, mitigating the key issue of low sensitivity. The GRS provides complementary information to the serum test and correctly reclassifies 20–37% of antibody-negative cases, potentially sparing the need for a kidney biopsy in this large subgroup of patients. In the clinical settings where neither a serum test nor a kidney biopsy is possible, the GRS itself can establish a diagnosis of MN with 99% specificity in 13–15% of cases. The practical advantage of this approach is that the GRS can be readily determined at any time after birth and, unlike the serum test, it does not fluctuate with time or in relationship to the disease onset, activity, or treatment. One important limitation, however, is that genetic effects may be population-specific and may not be generalizable to populations not represented in our GWAS. The performance of the GRS is remarkably consistent in our discovery and validation cohorts, including African-Americans, but it appears to be lower in self-reported Latino/Hispanics. Therefore, future efforts extending GWAS for MN to more diverse populations will be important.

In summary, we described a highly unusual genetic architecture of MN, including large effect sizes for a small number of common alleles and a strong evidence for ethnicity-specific genetic interactions. These insights enabled formulation a powerful genetic disease predictor that provides means to enhance a non-invasive diagnosis of MN, and can be especially useful in the settings where kidney biopsy represents too great of a risk or is not readily available.

Methods

Study design overview. We performed a genome-wide meta-analysis of eight discovery cohorts of East Asian and European ancestry (2956 cases and 7799 controls), all genotyped with high resolution arrays and imputed to ~7 million common high-quality markers using ancestry-matched reference panels. The top signals from the meta-analysis ($P < 5 \times 10^{-5}$) were typed in the additional East Asian replication cohort of 826 cases and 1239 controls. Subsequently, all cohorts (3782 cases and 9038 controls) were analyzed jointly to define genome-wide significant signals. All subjects provided informed consent to participate in genetic studies, and the Institutional Review Board of Columbia University as well as local ethics review committees for each of the individual cohorts approved our study protocol. The individual cohorts, genotyping methods, and quality control analyses are described in the Supplementary Methods.

Primary association analyses and genome-wide meta-analyses. Within each cohort, primary association scans were performed for markers that were common (MAF > 0.01) and imputed at high quality ($r^2 > 0.8$) using logistic regression under additive coding of dosage genotypes, and with adjustment for cohort-specific significant principal components (PCs) of ancestry. To quantify potential inflation of type I error due to stratification or technical artifacts, we estimated genomic inflation factors[45] for each genome-wide scan after excluding *HLA* and *PLA2R* loci. No substantial inflation was noted in any individual scan (lambda consistently <1.05 for each individual cohort). Subsequently, a fixed effects meta-analysis was performed to combine the results of the eight discovery cohorts using METAL[46]. Genome-wide distributions of *P*-values were examined visually using quantile-quantile plots for each individual cohort as well as for the combined analysis. The final meta-analysis quantile-quantile plot showed no global departures from the expected null distribution (Supplementary Fig. 3), with the genomic inflation factor estimated at 1.03 for the overall meta-analysis. Suggestive signals were defined by *P*-value $< 5.0 \times 10^{-5}$. To declare genome-wide significance of a novel locus, we used the generally accepted *P*-value threshold of 5.0×10^{-8}.

Conditional analyses. To detect additional independent SNPs at genome-wide significant loci, we performed stepwise conditional analyses of each locus using logistic regression. This was done by including the genotype of conditioning SNP (s) under additive coding as covariate(s) in the outcome model. The conditional analyses were performed individually within each cohort and with adjustments for cohort-specific ancestry PCs. Subsequently, the conditioned summary statistics were combined across cohorts using fixed effects meta-analysis, similar to our primary association analyses.

Credible set analyses. For each of the three genome-wide significant non-HLA loci, we derived 99% credible sets using the trans-ethnic and ethnicity-specific meta-analysis results. First, we derived approximate Bayes factors from GWAS association statistics using Wakefield's formula, as implemented in the R package gtx[47]. Using CAVIAR software[48], we next calculated the posterior probability (PP) for each SNP driving the association signal at each locus. We assumed there was only a single causal variant at each locus, since no additional independent SNPs were detected on stepwise conditioning analyses. We derived both trans-ethnic as well as ethnicity-specific 99% credible sets based on ranking the variants by their PPs and adding the variants to the set until cumulative PP > 99% was reached for each region. The overlaps between ethnicity-specific and trans-ethnic analyses were visualized in Supplementary Fig. 5.

HLA imputation. Six discovery cohorts (Chinese, South Korean, Japanese, European-1, European-2, and Turkish) had primary genotype data available for HLA imputation and association testing. For each of these cohorts, we imputed classical HLA alleles at two- and four-digit resolution, as well as individual amino acid polymorphisms at class I (*HLA-A*, *-B*, and *-C*) and class II (*HLA-DQB1*, *HLA-DQA1* and *HLA-DRB1*) loci using SNP2HLAsoftware[49]. The European cohorts and the East Asian cohorts were imputed separately, using ethnicity-specific reference panels. For European reference, we used the pre-phased HLA reference dataset generated by the Type 1 Diabetes Genetics Consortium (T1DGC, 5,225 individuals)[49]. For our East Asian cohorts, we used the Pan-Asian HLA Reference Panel (268 individuals)[50]. For validation of classical HLA association results in East Asians, we built additional East Asian reference panel based on the MHC sequence data from Zhou et al. (10,689 Han Chinese)[10]. In the association analyses, we included only common HLA alleles (MAF > 0.01) that were imputed with high certainty ($R^2 > 0.8$).

Statistical framework for HLA association testing. Given that the frequency of HLA alleles can vary by ethnicity, we performed HLA association testing Europeans and East Asians separately. We used logistic regression models to test the additive effects of HLA allele dosages with adjustment for significant PCs of ancestry. For multi-allelic loci, we used the following logistic regression model:

$$\log(\text{odds}_i) = \beta_0 + \sum_{j=1}^{m-1} \beta_j x_{j,i} + \sum_{k=1}^{n} \beta_k P_{k,i} \quad (1)$$

where m indicates a total number of alleles at a specific multi-allelic locus, j indicates a specific allele being tested, and $x_{j,i}$ is the imputed dosage for allele j for individual i; β_0 represents the intercept and β_j represents the additive effect of an allele j; $P_{k,i}$ denotes the value for kth PC of individual i, n is the total number of significant PCs in the dataset, β_k is the effect size of principal component k. We compared log-likelihoods of two nested models: the full model containing the test locus and relevant covariates with the reduced model without the test locus, but with the same set of covariates. The deviance was defined as −2×log likelihood ratio, which follows a X^2-distribution with m − 1 degrees of freedom, from which we calculated P-values. In addition to multi-allelic tests, we also performed bi-allelic tests of association for all individual SNPs, classical HLA alleles, and individual amino acid residues in HLA molecules. All analyses were performed using dosage method under additive coding. Stepwise conditioning analyses across the HLA region were performed using both multi-allelic and bi-allelic coding of HLA variants. In each round of stepwise conditioning, we first included the most significant variant as the covariate in the logistic regression model. If additional independently associated markers are detected, they are included as covariates in our subsequent models. We repeated these analyses until no residual associations across the entire locus were observed.

Analysis of polymorphic amino-acid sites in HLA genes. To test the effects of individual amino acid substitution sites within the *HLA-DRB1* and *HLA-DQA1* genes, we applied a conditional haplotype analysis using fully phased haplotypes across the HLA region. We tested each single amino acid position by first identifying the m possible amino acid residues occurring at that position and then using m − 1 degrees of freedom test to derive P-values, with a single amino acid residue arbitrarily selected as a reference. For conditioning on individual amino acid sites, we used the following procedure: by adding a new amino acid position to the model, a total of k additional unique haplotypes were generated and tested over the null model (without a new amino acid site) using the likelihood ratio test with k degrees of freedom. If the new position was independently significant, we further updated the null model to include all unique haplotypes created by all amino acid residues at both positons to identify another independent position. The procedure was repeated until no statistically significant positions were observed.

Testing for pairwise epistasis. Multiplicative interactive effects were tested using logistic regression model; SNPs were coded under additive genotype coding (0, 1, 2), and interaction terms were defined as simple products of genotypes. To screen for interactions, we tested a lead SNP at each of the three non-HLA loci against each of the five independent HLA SNPs (three in Europeans and two in East Asians) resulting in a total of 15 independent tests. We additionally tested for all pairwise interactions between the three non-HLA loci shared between both ethnicities resulting in three additional tests. In each case, we used a likelihood ratio test comparing two nested models: the full model with both main effects and the interaction term to the reduced model with main effects only. Given a total of 18 pairwise interaction tests, we used a Bonferroni-corrected significance threshold of $0.05/18 = 2.8 \times 10^{-3}$. In secondary analyses, we explored if significant HLA risk haplotypes interact with the *PLA2R1* risk allele in the six cohorts with fully imputed classical HLA alleles. This included a total of 2759 East Asians (803 cases and 1956 controls) and 4507 Europeans (1880 cases and 2627 controls).

Functional annotations of GWAS loci. We used several different approaches to perform functional annotations of our significant loci. We first defined the region of each locus as +/−400 kb of the index SNP. Using ANNOVAR software[51], we identified functional variants within each region that were in strong linkage disequilibrium ($r^2 > 0.8$) with the top SNP, including all known coding, splicing, 3′UTR and 5′UTR variants (Supplementary Table 12). To assess for potential functional variation in non-coding regions, we used our recently proposed tissue-specific functional scoring method (FUN-LDA)[11]. Using FUN-LDA, we estimated the posterior probability for each variant in strong LD with the top SNP of being functional across 127 different tissues or cell types profiled by ENCODE and ROADMAP consortia (Supplementary Figs. 6 and 10). To interrogate candidate variants against kidney-specific chromatin landscape, we analyzed regulatory DNase-seq maps paired with RNA-seq gene expression profiles from primary outgrowth cultures of human glomeruli (composed mainly of podocytes and mesangial cells) and renal cortex cultures (composed mainly of tubular cells), as well as chromatin conformation (Hi-C) maps from freshly isolated human glomeruli (Supplementary Figs. 9 and 11)[14]. To test for eQTL effects in kidney tissue, we used gene expression data from the NEPTUNE study[13]. This dataset is comprised of whole genome DNA sequence data and genome-wide transcriptome data (Affymetrix 2.1 ST chips) performed on micro-dissected glomerular ($N = 136$) and tubulointerstitial ($N = 166$) tissue compartments from kidney biopsies of patients with nephrotic syndrome. For each locus, we tested the index SNP and its high LD SNPs ($r^2 > 0.8$). Testing for cis-eQTL effects involved all transcripts within 1-Mb region centred on each SNP using the additive linear regression with adjustments for age, sex, PEER factors and first 4 PCs of ancestry as described previously[13]. In addition to kidney tissue, all loci were similarly interrogated for eQTL effects in the GTEx database version 8. Because *NFKB1* and *IRF4* both encode transcription factors, we also explored their potential binding in close proximity of each other or *PLA2R1* gene. Based on the Chip-seq data for all five subunits of NFκB complex in immortalized lymphocytes (GSE55105)[29], we found that the top SNPs at *PLA2R1* and *IRF4* loci intersect potential NFκB complex binding site (Supplementary Figs. 8 and 13). In addition, rs230492, a variant in strong LD (r2=0.94) with the top SNP at the *NFKB1* locus intersects a potential IRF4 binding site based on the Chip-seq data of IRF4 (GEO: GSM803390).

Pleiotropy analysis. We used the latest GWAS catalogue data to perform systematic cross-annotation of our top risk alleles against all other published GWAS findings. We first identified all genome-wide significant SNPs ($P < 5 \times 10^{-8}$) reported in the catalogue that resided within the genomic regions of association with MN. We assessed the extent of linkage disequilibrium (r^2) between these SNPs and the top MN risk alleles based on the combined European and East Asian sequence data from 1000 Genomes (phase 3). We next defined the directionality of pleiotropic effects as either concordant or opposed in relationship to the MN risk alleles. In addition, we queried each qualifying SNP from the catalogue against our genome-wide summary statistics to extract the odds ratios and P-values for associations with MN. We defined overlapping susceptibility alleles if r^2 exceeded 0.2 (Supplementary Table 13). Lastly, we constructed a susceptibility overlap map that connects each of the MN loci to the previously associated GWAS traits and highlights associations with SNPs in high LD with the top MN signals (Fig. 5a). The map was visualized with Cytoscape v.3.6 software.

Podocyte gene annotations of GWAS loci. To search for potential novel podocyte antigens encoded by our suggestive loci, we interrogated each locus against a podocyte-specific gene list predicted with in silico "nanodissection" approach (Supplementary Table 14). This computational approach used Affymetrix gene expression data from micro-dissected glomerular and tubulo-interstitial compartments of 452 renal biopsies[52]. In addition, we cross-annotated our positional candidates against the list of native podocyte proteins discovered by proteomic profiling of mouse podocytes[53]. All suggestive loci were additionally tested for multiplicative interactions with classical alleles, and the top most significant interactions were summarized in Supplementary Table 14. Lastly, we annotated all positional candidate genes using manual PubMed literature searches to prioritize genes with a previously established role in podocyte biology.

GR analysis. To estimate the cumulative effect of independently significant GWAS loci, we used a GRS approach. We first identified SNPs with independent contributions to MN risk at each locus using stepwise conditional analyses. Next, we tested for multiplicative interactions among those independently associated SNPs. We then built a logistic regression model including all independent SNPs along with their significant interaction terms to derive mutually adjusted effect sizes (Supplementary Table 15). Because we observed ethnicity-specific signals at the HLA locus, we generated ethnicity-specific for East Asians and Europeans separately. The ethnicity-specific risk scores were defined as a weighted sum of independent risk alleles and their significant interaction terms, weighted by their mutually adjusted effect sizes. The GRS was standardized using a Z-score transformation based on the mean and standard deviation for the distribution of ethnically matched controls, so that the standardized GRS was reflective of the distance between the raw score and the control mean in units of standard deviation. The final formulation of the GRS is as follows:

$$\begin{aligned}\text{East Asian GRS}_i = [&0.69173 \times d(rs9269027 - A)_i + 1.23685 \times d(rs1974461 - T)_i + 0.36687 \\ &\times d(rs6707458 - G)_i + 0.25098 \times d(rs230540 - C)_i + 0.39127 \\ &\times d(rs9405192 - G)_i + 0.48798 \times d(rs9269027 - A)_i \times d(rs6707458 - G)_i \\ &- \text{East Asian Control Mean GRS}]/\ \text{East Asian Control SD GRS}\end{aligned} \quad (2)$$

where GRS_i = genetic risk score for individual i, d_i = dosage of risk allele (from 0 to 2) for individual i, East Asian Control Mean GRS = 1.6804, East Asian Control SD GRS = 1.003

$$\begin{aligned}\text{European GRS}_i = [&0.34945 \times d(rs9271541 - C)_i + 0.67919 \times d(rs9265949 - T)_i \\ &+ 0.30707 \times d(rs2858309 - C)_i + 0.34601 \times d(rs6707458 - G)_i \\ &+ 0.17450 \times d(rs230540 - C)_i + 0.18343 \times d(rs9405192 - G)_i \\ &+ 0.33782 \times d(rs9271541 - C)_i \times d(rs6707458 - G)_i \\ &- \text{European Control Mean GRS}]/\text{European Control SD GRS}\end{aligned} \quad (3)$$

where GRS_i = genetic risk score for individual i, d_i = dosage of risk allele (from 0 to 2) for individual i, European GRS control mean = 1.5089, European GRS control SD = 0.8202.

The percentage of the total variance in disease risk explained was estimated using Nagelkerke's pseudo R^2 from the logistic regression model with the standardized GRS as a predictor and case-control status as an outcome. The performances of the ethnicity-specific GRS were estimated by the AUROC. We performed detailed GRS cut-off analyses in the discovery cohorts by selecting cut-off points on the ROC curve that provide specificities in the range from 95 to 100%. For each cut-off point, we calculate sensitivity, specificity, positive likelihood ratio (LR+), and negative likelihood ratio (LR-). All GRS analyses were implemented in R version 3.3.2.

CRS formulation. The CRS was formulated as a weighted sum of the GRS and serum anti-PLA2R antibody levels in U/mL. The weight was determined using logistic regression model, with standardized ethnicity-specific GRS (formulated as described above) and natural log-transformed anti-PLA2R antibody levels as two predictors and case-control status as an outcome. This resulted in the following model:

$$Y_i = \beta_0 + \beta_1 \mathrm{GRS}_i + \beta_2 \ln(\alpha\mathrm{PLA2R}_i + 0.001) \quad (4)$$

where GRS_i indicates the Z-transformed ethnicity-specific GRS for individual i, β_0 indicates the intercept, β_1 indicates the effect size of the GRS estimated based on discovery cohorts; $\alpha\mathrm{PLA2R}_i$ is the serum level of PLA2R antibodies in U/mL; the constant 0.001 is added to enable log transformation of undetectable (zero) levels; β_2 represents the effect size for anti-PLA2R antibody positivity estimated based on discovery cohorts. The weight for the $\ln(\alpha\mathrm{PLA2R}_i + 0.001)$ term was then defined as follows:

$$\text{Weight} = \frac{\beta_2}{\beta_1} \quad (5)$$

Consequently, the weights for antibody levels were calculated for East Asian and European separately, which resulted in the following Crude CRS formulation:

$$\text{Crude CRS}_i = \begin{cases} \mathrm{GRS}_i + 0.4829 \times \ln(\alpha\mathrm{PLA2R}_i + 0.001), \text{if } i \in \text{European} \\ \mathrm{GRS}_i + 1.7712 \times \ln(\alpha\mathrm{PLA2R}_i + 0.001), \text{if } i \in \text{East Asian} \end{cases} \quad (6)$$

In the final step, the Crude CRS was Z-transformed using the mean and standard deviation for ethnicity-matched healthy controls:

$$\mathrm{CRS}_i = \begin{cases} (\text{Crude CRS}_i - \text{European Control Mean CRS})/(\text{European Control SD}), \text{if } i \in \text{European} \\ (\text{Crude CRS}_i - \text{E.Asian Control Mean CRS})/(\text{E.Asian Control SD}), \text{if } i \in \text{East Asian} \end{cases} \quad (7)$$

where European Control Mean CRS = −1.4982, European Control SD = 1.4354, E. Asian Control Mean CRS = 0.3724, E. Asian Control SD = 2.7503.

The CRS performance was estimated by the area under the receiver operating curve (AUROC). Similar to GRS, we explored different CRS cut-offs that maximized specificity in the range 95 to 100%. For each cut-off, we calculated sensitivity, specificity, positive likelihood ratio (LR+), and negative likelihood ratio (LR−). The integrated discrimination improvement (IDI) and net reclassification improvement (NRI)[54] were calculated comparing the CRS test to the serum PLA2R antibody test alone. All CRS analyses were implemented in R version 3.3.2 (CRAN).

European GWAS validation cohorts. For the purpose of GRS validation, we utilized three previously published external GWAS cohorts of European ancestry: the UK, French, and Dutch Validation Cohorts[6]. These cohorts were composed of biopsy-documented cases of primary MN and ethnicity-matched healthy controls, totalling 2,887 individuals (550 cases and 2337 controls, see Supplementary Methods for details). For all three cohorts, we obtained primary genotype data after quality control analysis as published previously[6] and performed phasing and imputation using Eagle v.2.3 and Minimac3 with European populations of Phase 3 1000 Genome Project as a reference. The cases and controls were imputed jointly. The GRS for each individual was determined using genotype dosages for the SNPs included in the score. The performance of GRS was then analyzed individually in each cohort, and in all cohorts combined (Supplementary Tables 17–19).

NEPTUNE GRS validation cohorts. The Nephrotic Syndrome Study Network (NEPTUNE) is a prospective, longitudinal cohort recruiting participants with nephrotic syndrome at the time of first kidney biopsy[8]; primary disease diagnoses of MN, FSGS, MCD, and IgAN were determined by a central pathology review. All NEPTUNE participants underwent low-depth whole genome sequencing as described in the Supplementary Methods. The GRS was successfully determined for $N = 475$ NEPTUNE participants. This included 89 cases with the diagnosis of primary MN and 386 disease controls, including 184 with FSGS, 164 with MCD, and 38 with IgAN. Our pre-specified primary GRS validation involved 180 NEPTUNE participants of European ancestry (46 cases and 134 disease controls, Supplementary Tables 17–18). In secondary analyses, we extended our validation studies to the entire multiethnic cohort of 475 NEPTUNE participants (Supplementary Tables 20–21), including subgroup analyses of 133 individuals of African American ancestry (18 cases and 115 disease controls) and 94 individuals of Hispanic/Latino American ancestry (18 cases and 76 disease controls). The numbers of NEPTUNE participants in other ancestral groups were too small for a meaningful analysis.

PLA2R antibody testing. In total, we determined serum antibody levels in $N = 2331$ study participants with genetic data ($N = 1488$ cases, $N = 300$ healthy controls, and $N = 543$ disease controls) across all cohorts and ethnicities. The ancestry-matched diseased controls were recruited among patients commonly presenting with nephrotic syndrome, including FSGS, MCD, IgAN. For all MN cases and disease controls, serum samples were obtained near or at the time of kidney biopsy and any samples obtained more than six months after the biopsy were excluded from the analysis. All individuals that underwent antibody testing had matched genetic data, enabling derivation and diagnostic testing of the CRS.

We performed a standardized measurement of serum anti-PLA2R Ab levels using the anti-PLA2R ELISA (IgG) test kit (EUROIMMUN Medizinische Labordiagnostika AG), which employs the indirect ELISA methodology. The kit includes a 96-well microplate pre-coated with PLA2R, 5 calibrators (2, 20, 100, 500, and 1500 U/mL respectively), positive and negative control samples, peroxidase-labelled anti-human IgG (rabbit) enzyme conjugate, kit specific sample and wash buffers, Chromogen/substrate solution (TMB/H_2O_2), and stop solution. The assay was run as per the protocol included with the kit. A 5-point calibrated analysis was used to calculate the results for each assay performed. A standard curve was generated based on the spectrophotometric reading of the five calibrators included on each microplate. As recommended by EUROIMMUN, the sample was called positive if antibody level was ≥20 U/mL.

In the discovery stage, we analyzed sera for a total of $N = 459$ East Asian and $N = 1034$ European participants. The European cohorts included 810 cases with MN, 99 healthy controls, and 125 disease controls (37 FSGS and 88 IgAN). The East Asian cohorts included 304 cases with MN, 56 healthy controls, and 99 disease controls (52 FSGS and 47 IgAN). The disease controls were not included in the GWAS discovery analysis, but were added to test for disease specificity of the serologic test. We note that one East Asian disease control with a clinical diagnosis of IgAN tested anti-PLA2R antibody positive at high titer; follow-up pathology review found evidence of previously unrecognized sub-epithelial deposits diagnostic of MN in addition to IgA-dominant mesangial deposits; this individual was subsequently removed from the analysis.

In the validation phase, we analyzed sera for a total of $N = 540$ individuals including European case-control cohorts ($N = 248$ cases and $N = 145$ healthy controls) and a total of 147 European NEPTUNE participants ($N = 36$ cases and $N = 111$ disease controls). In secondary analysis, we extended testing to additional 180 NEPTUNE participants of non-European ancestry ($N = 28$ cases and $N = 152$ controls). This included 103 NEPTUNE participants of African American ancestry (16 cases and 87 disease controls) and 77 participants of Hispanic/Latino ancestry (12 cases and 65 disease controls). These numbers are smaller compared to the GRS validation cohorts, since not all NEPTUNE participants had sera sampled within 6 months of biopsy.

Testing for phenotypic correlations of the GRS. Using extensive clinical information available for our discovery cohorts, we investigated correlations between GRS and clinical traits from the time of kidney biopsy including age at diagnosis, 24-h proteinuria (P24), estimated glomerular filtration rate (eGFR), serum albumin (Alb) and serum anti-PLA2R1 antibody level (Supplementary Table 16). The eGFR was estimated based on serum creatinine level using the Modification of Diet in Renal Disease (MDRD) equation[55]. The proteinuria was quantified using spot urine protein-to-creatinine ratios. The values of proteinuria and eGFR were normalized by natural log-transformation. The serum albumin and anti-PLA2R1 antibody levels also required natural log-transformation. For each quantitative trait, we built a linear regression model with GRS as a predictor and each corresponding trait as an outcome. For dichotomous traits (anti-PLA2R seropositivity or presence of nephrotic range proteinuria), we used logistic regression with a GRS as a predictor and each binary trait as an outcome. The association analysis for age at diagnosis (biopsy) was performed before and after adjustment for sex and ancestry. The tests of proteinuria, eGFR, serum albumin and serum anti-PLA2R1 antibody levels were carried out before and after controlling for age, sex and ancestry. Statistical analyses were implemented in R version 3.3.2.

SNP-based heritability. We estimated the SNP-based heritability of MN in East Asians and Europeans using the GCTA-GREML algorithm[30,56]. For this analysis, we included three European cohorts (European discovery 1, European discovery 2 and Turkish discovery) and three East Asian cohorts (Chinese, Korean and Japanese discovery) that had primary genotype data available for joint heritability analysis. We first estimated pairwise genetic relationship matrix between all individuals using autosomal SNPs. With the GCTA software[57], we next estimated the disease variance explained by all autosomal SNPs. We transformed the estimate assuming an underlying liability scale and disease prevalence of 0.001. We derived a standard error (SE) and 95% confidence interval for each estimate.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All genome-wide summary statistics, including those presented in Fig. 1, are freely available for download on our lab website: www.columbiamedicine.org/divisions/kiryluk/resources.php. The calculations of genetic risk score (GRS) and combined risk score (CRS) are implemented in the form of an online risk calculator, which is also freely available on our lab website. The PAGE consortium control genotype data is available on dbGAP under accession number phs000356.v2.p1. Primary genotype data for the European-1 discovery cohort is available under dbGAP accession number phs001984.v1.p1. Our IRB determined that the use of this dataset is restricted to genetic studies of kidney disease. Because of consent restrictions and/or country-specific privacy laws, we are unable to share primary genotype data on dbGAP for other cohorts. All data and summary statistics are available from the corresponding authors upon reasonable request.

Received: 28 August 2019; Accepted: 3 March 2020;
Published online: 30 March 2020

References

1. Glassock, R. J. Diagnosis and natural course of membranous nephropathy. *Semin Nephrol.* **23**, 324–332 (2003).
2. Debiec, H. et al. Antenatal membranous glomerulonephritis due to anti-neutral endopeptidase antibodies. *N. Engl. J. Med.* **346**, 2053–2060 (2002).
3. Beck, L. H. Jr. et al. M-type phospholipase A2 receptor as target antigen in idiopathic membranous nephropathy. *N. Engl. J. Med.* **361**, 11–21 (2009).
4. Glassock, R. J. The pathogenesis of membranous nephropathy: evolution and revolution. *Curr. Opin. Nephrol. Hypertens.* **21**, 235–242 (2012).
5. Tomas, N. M. et al. Thrombospondin type-1 domain-containing 7A in idiopathic membranous nephropathy. *N. Engl. J. Med.* **371**, 2277–2287 (2014).
6. Stanescu, H. C. et al. Risk HLA-DQA1 and PLA(2)R1 alleles in idiopathic membranous nephropathy. *N. Engl. J. Med.* **364**, 616–626 (2011).
7. Wunnenburger, S. et al. Associations between genetic risk variants for kidney diseases and kidney disease etiology. *Sci. Rep.* **7**, 13944 (2017).
8. Gadegbeku, C. A. et al. Design of the Nephrotic Syndrome Study Network (NEPTUNE) to evaluate primary glomerular nephropathy by a multidisciplinary approach. *Kidney Int.* **83**, 749–756 (2013).
9. Cui, Z. et al. MHC class II risk alleles and amino acid residues in idiopathic membranous nephropathy. *J. Am. Soc. Nephrol.* **28**, 1651–1664 (2017).
10. Zhou, F. et al. Deep sequencing of the MHC region in the Chinese population contributes to studies of complex disease. *Nat. Genet.* **48**, 740–746 (2016).
11. Backenroth, D. et al. FUN-LDA: a latent dirichlet allocation model for predicting tissue-specific functional effects of noncoding variation: methods and applications. *Am. J. Hum. Genet* **102**, 920–942 (2018).
12. Consortium, G. T. et al. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
13. Gillies, C. E. et al. An eQTL landscape of kidney tissue in human nephrotic syndrome. *Am. J. Hum. Genet.* **103**, 232–244 (2018).
14. Sieber, K. B. et al. Integrated functional genomic analysis enables annotation of kidney genome-wide association study loci. *J. Am. Soc. Nephrol.* **30**, 421–441 (2019).
15. Westra, H. J. et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet* **45**, 1238–1243 (2013).
16. Raj, T. et al. Polarization of the effects of autoimmune and neurodegenerative risk alleles in leukocytes. *Science* **344**, 519–523 (2014).
17. Astle, W. J. et al. The allelic landscape of human blood cell trait variation and links to common complex disease. *Cell* **167**, 1415–1429 e19 (2016).
18. Jostins, L. et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119–124 (2012).
19. Liu, J. Z. et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet* **47**, 979–986 (2015).
20. Mells, G. F. et al. Genome-wide association study identifies 12 new susceptibility loci for primary biliary cirrhosis. *Nat. Genet* **43**, 329–332 (2011).
21. Cordell, H. J. et al. International genome-wide meta-analysis identifies new primary biliary cirrhosis risk loci and targetable pathogenic pathways. *Nat. Commun.* **6**, 8019 (2015).
22. Wu, H. et al. Comparative analysis and refinement of human PSC-derived kidney organoid differentiation with single-cell transcriptomics. *Cell Stem Cell* **23**, 869–881.e8 (2018).
23. Pattaro, C. et al. Genetic associations at 53 loci highlight cell types and biological pathways relevant for kidney function. *Nat. Commun.* **7**, 10023 (2016).
24. Wuttke, M. et al. A catalog of genetic loci associated with kidney function from analyses of a million individuals. *Nat. Genet.* **51**, 957–972 (2019).
25. Grumont, R. J. & Gerondakis, S. Rel induces interferon regulatory factor 4 (IRF-4) expression in lymphocytes: modulation of interferon-regulated gene expression by rel/nuclear factor kappaB. *J. Exp. Med.* **191**, 1281–1292 (2000).
26. Saito, M. et al. A signaling pathway mediating downregulation of BCL6 in germinal center B cells is blocked by BCL6 gene alterations in B cell lymphoma. *Cancer Cell* **12**, 280–292 (2007).
27. Boddicker, R. L. et al. The oncogenic transcription factor IRF4 is regulated by a novel CD30/NF-kappaB positive feedback loop in peripheral T-cell lymphoma. *Blood* **125**, 3118–3127 (2015).
28. Lake, B. B. et al. A single-nucleus RNA-sequencing pipeline to decipher the molecular anatomy and pathophysiology of human kidneys. *Nat. Commun.* **10**, 2832 (2019).
29. Zhao, B. et al. The NF-kappaB genomic landscape in lymphoblastoid B cells. *Cell Rep.* **8**, 1595–1606 (2014).
30. Lee, S. H., Wray, N. R., Goddard, M. E. & Visscher, P. M. Estimating missing heritability for disease from genome-wide association studies. *Am. J. Hum. Genet.* **88**, 294–305 (2011).
31. Bobart, S. A. et al. Noninvasive diagnosis of primary membranous nephropathy using phospholipase A2 receptor antibodies. *Kidney Int* **95**, 429–438 (2019).
32. Schmid, H. et al. Modular activation of nuclear factor-kappaB transcriptional programs in human diabetic nephropathy. *Diabetes* **55**, 2993–3003 (2006).
33. Atreya, I., Atreya, R. & Neurath, M. F. NF-kappaB in inflammatory bowel disease. *J. Intern. Med.* **263**, 591–596 (2008).
34. Schottelius, A. J. & Baldwin, A. S. Jr A role for transcription factor NF-kappa B in intestinal inflammation. *Int J. Colorectal Dis.* **14**, 18–28 (1999).
35. Mezzano, S. A. et al. Tubular NF-kappaB and AP-1 activation in human proteinuric renal disease. *Kidney Int.* **60**, 1366–1377 (2001).
36. Mudge, S. J., Paizis, K., Auwardt, R. B., Thomas, R. J. & Power, D. A. Activation of nuclear factor-kappa B by podocytes in the autologous phase of passive Heymann nephritis. *Kidney Int.* **59**, 923–931 (2001).
37. Liu, S. et al. Urinary messenger RNA of the receptor activator of NF-kappaB could be used to differentiate between minimal change disease and membranous nephropathy. *Biomarkers* **19**, 597–603 (2014).
38. Bonder, M. J. et al. Disease variants alter transcription factor levels and methylation of their binding sites. *Nat. Genet.* **49**, 131–138 (2017).
39. Hu, X. et al. Additive and interaction effects at three amino acid positions in HLA-DQ and HLA-DR molecules drive type 1 diabetes risk. *Nat. Genet.* **47**, 898–905 (2015).
40. Raychaudhuri, S. et al. Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis. *Nat. Genet.* **44**, 291–296 (2012).
41. Xu, X. et al. Molecular insights into genome-wide association studies of chronic kidney disease-defining traits. *Nat. Commun.* **9**, 4800 (2018).
42. Sekula, P. et al. Genetic risk variants for membranous nephropathy: extension of and association with other chronic kidney disease aetiologies. *Nephrol. Dial. Transpl.* **32**, 325–332 (2017).
43. Nevalainen, T. J., Graham, G. G. & Scott, K. F. Antibacterial actions of secreted phospholipases A2. Review. *Biochim Biophys. Acta* **1781**, 1–9 (2008).
44. Morri, H., Ozaki, M. & Watanabe, Y. 5'-flanking region surrounding a human cytosolic phospholipase A2 gene. *Biochem. Biophys. Res. Commun.* **205**, 6–11 (1994).
45. Devlin, B., Roeder, K. & Bacanu, S. A. Unbiased methods for population-based association studies. *Genet. Epidemiol.* **21**, 273–284 (2001).
46. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
47. Wakefield, J. Bayes factors for genome-wide association studies: comparison with P-values. *Genet. Epidemiol.* **33**, 79–86 (2009).
48. Hormozdiari, F., Kostem, E., Kang, E. Y., Pasaniuc, B. & Eskin, E. Identifying causal variants at loci with multiple signals of association. *Genetics* **198**, 497–508 (2014).
49. Jia, X. et al. Imputing amino acid polymorphisms in human leukocyte antigens. *PLoS ONE* **8**, e64683 (2013).
50. Pillai, N. E. et al. Predicting HLA alleles from high-resolution SNP data in three Southeast Asian populations. *Hum. Mol. Genet.* **23**, 4443–4451 (2014).
51. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
52. Ju, W. et al. Defining cell-type specificity at the transcriptional level in human disease. *Genome Res.* **23**, 1862–1873 (2013).
53. Rinschen, M. M. et al. A multi-layered quantitative in vivo expression atlas of the podocyte unravels kidney disease candidate genes. *Cell Rep.* **23**, 2495–2508 (2018).
54. Steyerberg, E. W. et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* **21**, 128–138 (2010).
55. Levey, A. S. et al. Using standardized serum creatinine values in the modification of diet in renal disease study equation for estimating glomerular filtration rate. *Ann. Intern. Med.* **145**, 247–254 (2006).

56. Yang, J. et al. Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).
57. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).

Acknowledgements

We are grateful to all study participants across multiple nephrology centres worldwide for their contributions to this work. This work was supported by the following institutions, grants and funding agencies in the US: Columbia University, Columbia Glomerular Center, National Institute for Diabetes and Digestive Kidney Diseases (NIDDK) grants RC2-DK116690 (K.K., M.K.), R01-DK105124 (K.K.), R01-DK097053 (L.H.B., M.K.), R01-DK108805 (M.G.S.), and National Institute on Minority Health and Health Disparities (NIMHD) grant R01-MD009223 (K.K., A.G.G., A.B.). M.G.S. is additionally supported by the Charles Woodson Clinical Research Fund. The Nephrotic Syndrome Study Network Consortium (NEPTUNE), U54-DK-083912, is a part of the National Institutes of Health (NIH) Rare Disease Clinical Research Network (RDCRN), supported through a collaboration between the Office of Rare Diseases Research, National Center for Advancing Translational Sciences (NCATS) and the National Institute of Diabetes, Digestive, and Kidney Diseases (NIDDK). Additional funding and/or programmatic support for this project has also been provided by the University of Michigan, the NephCure Kidney International and the Halpin Foundation. The recruitment and analysis of the Chinese cohorts were supported by the National Key Research and Development Program of China (2016YFC0904100), Natural Science Foundation of China to the Innovation Research Group (81621092), National Natural Science Foundation of China (No. 81870460, 81570598), Science and Technology Innovation Action Plan of Shanghai Science and Technology Committee (No.17441902200), Shanghai Municipal Education Commission, Gaofeng, Clinical Medicine Grant (No.20152207), Shanghai Jiao Tong University School of Medicine, Multi-Center Clinical Research Project (No.DLY201510), International Cooperation and Exchange Projects of Shanghai Science and Technology Committee (No.14430721000), the Outstanding Young Scholar Award for Zhao Cui (No.81622009), and Shanghai Health and Family Planning Committee Hundred Talents Program for Jingyuan Xie (No.2018BR37). The recruitment of the Korean cohort was supported by the Seoul National University Hospital Human Biobank, a member of the National Biobank of Korea, financed by the Ministry of Health and Welfare, Republic of Korea. P.B. and P.H. acknowledge financial support from MRC project "Autoimmunity in Membranous Nephropathy", grant MR/J010847/1 which funded the sample collection from MN patients across the UK. P.B., P.H. and S.H. acknowledge support from Manchester Academic Health Science Centre (MAHSC 186/200), the Greater Manchester Local Clinical Research Network and Kidneys for Life Charity for supporting research in MN in Manchester. We are grateful to the MENTOR study (clinical trials no. NCT01180036), for contributing blood samples of trial participants. The UK cohort was supported in part by grants from the David and Elaine Potter Charitable Foundation (to S.H.P. and R.K.), St Peter's Trust for Kidney, Bladder and Prostate Research (to D.B., H.C.S., S.H.P. and R.K.), Kids Kidney Research UK and Kidney Research UK (to D.B. and R.K.). The Italian cohorts were supported by the Italian Ministry of Health grant GR-2011-02350438 (G.Z., S.G.) and the Department of Excellence Grant 2018–2022 funded by the Italian Ministry of Education for the Department of Medical Sciences of the University of Turin (A.A.). The recruitment of Polish cases was sponsored by the Polish Kidney Genetics Network (POLYGENES), a collaborative effort between Columbia University and Poznań University of Medical Sciences, Poland. The full list of POLYGENES collaborators can be found in the Supplementary Materials. The GCKD (German Chronic Kidney Disease) study was funded by grants from the German Ministry of Education and Research (BMBF, No. 01ER0804) and the KfH Foundation for Preventive Medicine, with genotyping supported by Bayer Pharma AG. The list of GCKD investigators can be found in the Supplementary Materials. The work of M.W. and A.K. was funded by the CRC 1140 Initiative and by KO 3598/3–1 and CRC 992 (A.K.) of the German Research Foundation. The work of E.H. and R.A.K.S. was funded by the CRC 1192 from the German Research Foundation (Projects B1 and C1). P.R. is a recipient of European Research Council ERC-2012-ADG_20120314 grant 322947, 7th Framework Programme of the European Community contract 2012–305608 (European Consortium for High-Throughput Research in Rare Kidney Diseases), and the National Research Agency grant MNaims (ANR-17- CE17-0012-01). The Dutch studies were supported by grants from the Dutch Kidney Foundation to JMH and JFW (Nierstichting Nederland grant OW08 and grant KJPB11.021). We would like to thank the Population Architecture Using Genomics and Epidemiology (PAGE) consortium, funded by the National Human Genome Research Institute (NHGRI) with co-funding from the NIMHD, for providing population controls for this study. For full acknowledgment of the PAGE consortium, please see Supplementary Materials. The funding sources were not involved in the study design, collection, analysis, and interpretation of data, writing of the report, or in the decision to submit the paper for publication.

Author contributions

K.K., N.C. and R.K. conceived the study and made the decision to publish the findings; K. K. designed the study and provided overall supervision of the project; K.K. and L. Liu wrote the initial draft of the manuscript; J. Xie coordinated recruitment and genotyping of the Chinese discovery and replication cohorts; N.M. and J. Xie performed quality control, imputation and GWAS association analyses for Asian discovery cohorts; L. Liu performed quality control, imputation and GWAS association analyses for European discovery and validation cohorts, and performed final statistical and bioinformatics analyses, including GWAS meta-analyses, fine-mapping, HLA analyses, functional annotations, risk score analyses, clinical correlation analyses, and generated figures and tables; I.I.-L. consulted on the statistics and functional annotation of GWAS loci. A.G.G. consulted on the study design and critically reviewed the manuscript. J. Xie, X.Y., O.B., L.H.B., H.D., E.H., R.A.K.S., P.B. and R.K. performed serum antibody studies. O.B., Y.L., J.Y.Z., P.K., N.M., R.J.R., M. Bodria, A. Khan, K. Mehl, and F.O. contributed to sample processing, biobanking, DNA extractions, quality control, plating, clinical and genetic data management. S.A. generated kidney compartment-specific maps of chromatin accessibility and interactions. J. Xie, N.C., M-H.Z., Z.C., L.H.R., W.W., Z.L., X.H., X.Y., D.Z., J. Xu, G.L. recruited, genotyped, and clinically characterized the Chinese Discovery and Chinese Replication cohorts. A.P., C.S., M.Z., and F.C. genotyped, imputed, and analyzed GWAS data for the Sardinian Discovery cohort. M.W., K-U.E., and A. Köttgen coordinated the GCKD Study, genotyped, imputed and analyzed GWAS data for the GCKD Discovery cohort. H.S. recruited and clinically characterized the Japanese Discovery cohort. H.L., J.P., B.L.C., Y.S.K., and D.K.K. recruited and clinically characterized the Korean Discovery cohort. Y.C., S.O., A.Y., N.S., H.A., M. Koc, T.B., G.K., S.U.A., and M.S.S. recruited and clinically characterized the Turkish Discovery cohort. M.M., P.A.C., A.S.B., G.B.A., S.S.-C., M. Bodria, F.Z., A.P.-P., M.D., K. Mucha, B.M., B.F., L.P., I.H., E.A., J.B., L.M., B.V., D. Santoro, M. Bonomini, F. Londrino, L.G., J.R., V.T., C.I., S.S., D. Spotti, C.M., P.M., M.G., D.R., S. Granata, G.Z., F. Lugani, G.M.G., I.P., L.A., B. Sprangers, D.C.C., F.C.F., A.A., and F.S. recruited and clinically characterized patients for the European-1 Discovery Cohort. P.H., S.H., S. Gupta, C.C., S.D., N.I., R.J.P., J.C., S.P., D.B., H.C.S., N.A., E.H., R.A.K.S., P.B. and R.K. participated in the recruitment, clinical characterization and generation of genotype data for the European Discovery-2 and the UK Validation cohorts. B. Stengel, H.D., and P.R. contributed the French Validation Cohort. J.M.H., M.J.C., L.A.K., and J.F.M.W. contributed the Dutch Validation Cohort. M.G.S., L.H.M., L.H.B. and M. Kretzler contributed the NEPTUNE validation cohort. R.J.F.L. and E.E.K. contributed the control genotype data for the PAGE cohort. All authors have read and approved the final version of manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41467-020-15383-w.

Correspondence and requests for materials should be addressed to J.X., R.K. or K.K.

Peer review information *Nature Communications* thanks Yang Luo and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Jingyuan Xie[1,77✉], Lili Liu[2,77], Nikol Mladkova[2,77], Yifu Li[2], Hong Ren[1], Weiming Wang[1], Zhao Cui[3,4,5], Li Lin[1], Xiaofan Hu[1], Xialian Yu[1], Jing Xu[1], Gang Liu[3,4,5], Yasar Caliskan[6], Carlo Sidore[7], Olivia Balderes[2], Raphael J. Rosen[2], Monica Bodria[2,8], Francesca Zanoni[2,9], Jun Y. Zhang[2], Priya Krithivasan[2], Karla Mehl[2], Maddalena Marasa[2], Atlas Khan[2], Fatih Ozay[2], Pietro A. Canetta[2], Andrew S. Bomback[2], Gerald B. Appel[2], Simone Sanna-Cherchi[2], Matthew G. Sampson[10], Laura H. Mariani[11,12], Agnieszka Perkowska-Ptasinska[13], Magdalena Durlik[13], Krzysztof Mucha[14,15], Barbara Moszczuk[14], Bartosz Foroncewicz[14], Leszek Pączek[14,15], Ireneusz Habura[16], Elisabet Ars[17], Jose Ballarin[17], Laila-Yasmin Mani[18], Bruno Vogt[18], Savas Ozturk[19], Abdülmecit Yildiz[20], Nurhan Seyahi[21], Hakki Arikan[22], Mehmet Koc[22], Taner Basturk[23], Gonca Karahan[24], Sebahat Usta Akgul[24], Mehmet Sukru Sever[6], Dan Zhang[25], Domenico Santoro[26], Mario Bonomini[27], Francesco Londrino[28], Loreto Gesualdo[29], Jana Reiterova[30], Vladimir Tesar[30], Claudia Izzi[31,32], Silvana Savoldi[33], Donatella Spotti[34], Carmelita Marcantoni[35], Piergiorgio Messa[9], Marco Galliani[36], Dario Roccatello[37], Simona Granata[38], Gianluigi Zaza[38], Francesca Lugani[39], GianMarco Ghiggeri[39], Isabella Pisani[8], Landino Allegri[8], Ben Sprangers[40,41], Jin-Ho Park[42], BeLong Cho[42,43], Yon Su Kim[44,45], Dong Ki Kim[45,46], Hitoshi Suzuki[47], Antonio Amoroso[48], Daniel C. Cattran[49], Fernando C. Fervenza[50], Antonello Pani[51], Patrick Hamilton[52], Shelly Harris[52], Sanjana Gupta[53], Chris Cheshire[53], Stephanie Dufek[53], Naomi Issler[53], Ruth J. Pepper[53], John Connolly[53], Stephen Powis[53], Detlef Bockenhauer[53], Horia C. Stanescu[53], Neil Ashman[54], Ruth J.F. Loos[55,56,57], Eimear E. Kenny[55,58,59], Matthias Wuttke[60], Kai-Uwe Eckardt[61,62], Anna Köttgen[60], Julia M. Hofstra[63], Marieke J.H. Coenen[64], Lambertus A. Kiemeney[65], Shreeram Akilesh[66], Matthias Kretzler[11], Lawrence H. Beck[67], Benedicte Stengel[68,69], Hanna Debiec[70], Pierre Ronco[70,71], Jack F.M. Wetzels[63], Magdalena Zoledziewska[7], Francesco Cucca[7], Iuliana Ionita-Laza[72], Hajeong Lee[45,46], Elion Hoxha[73], Rolf A.K. Stahl[73], Paul Brenchley[74], Francesco Scolari[31,75], Ming-hui Zhao[3,4,5,76], Ali G. Gharavi[2], Robert Kleta[53,77✉], Nan Chen[1,77] & Krzysztof Kiryluk[2,77✉]

[1]Department of Nephrology, Institute of Nephrology, Shanghai Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China. [2]Department of Medicine, Division of Nephrology, Columbia University, College of Physicians & Surgeons, New York, USA. [3]Renal Division, Department of Medicine, Peking University First Hospital, Beijing, China. [4]Institute of Nephrology, Peking University, Beijing, China. [5]Key Laboratory of Renal Disease, Ministry of Health of China, and Key Laboratory of CKD Prevention and Treatment, Ministry of Education of China, Beijng, China. [6]Division of Nephrology, Department of Internal Medicine, Istanbul School of Medicine, Istanbul University, Istanbul, Turkey. [7]Istituto di Ricerca Genetica e Biomedica, Consiglio Nazionale delle Ricerche, Monserrato, Cagliari, Italy. [8]Department of Medicine and Surgery, University of Parma, Parma, Italy. [9]Nephrology Dialysis and Kidney Transplant Unit, Fundazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Università degli studi di Milano, Milan, Italy. [10]Department of Pediatrics-Nephrology, University of Michigan School of Medicine, Ann Arbor, MI, USA. [11]Division of Nephrology, Department of Medicine, University of Michigan, Ann Arbor, MI, USA. [12]Arbor Research Collaborative for Health, Ann Arbor, MI, USA. [13]Department of Transplantology, Nephrology and Internal Diseases, Medical University of Warsaw, Warsaw, Poland. [14]Department of Immunology, Transplantology and Internal Diseases, Medical University of Warsaw, Warsaw, Poland. [15]Institute of Biochemistry and Biophysics, Polish Academy of Sciences, Warsaw, Poland. [16]Department of Nephrology, University Hospital of Karol Marcinkowski in Zielona Góra, Zielona Góra, Poland. [17]Molecular Biology Laboratory and Nephrology Department, Fundació Puigvert, Instituto de Investigaciones Biomédicas Sant Pau, Universitat Autònoma de Barcelona, REDINREN, IISCIII, Barcelona, Spain. [18]Department of Nephrology and Hypertension, Bern University Hospital, University of Bern, Bern, Switzerland. [19]Nephrology Clinic, Haseki Training and Research Hospital, Istanbul, Turkey. [20]Department of Nephrology, Uludag University Faculty of Medicine, Bursa, Turkey. [21]Division of Nephrology, Department of Internal Medicine, Cerrahpasa Medical Faculty, Istanbul University - Cerrahpasa, Istanbul, Turkey. [22]Division of Nephrology, Department of Internal Medicine, Marmara University School of Medicine, Istanbul, Turkey. [23]Department of Nephrology, Sisli Hamidiye Etfal Training and Research Hospital, Istanbul, Turkey. [24]Department of Medical Biology, Istanbul School of Medicine, Istanbul University, Istanbul, Turkey. [25]Department of Nephrology, Xin Hua Hospital Affiliated to Shanghai Jiao Tong University School of Medicine, Shanghai, China. [26]Department of Cinical and Experimental Medicine, Unit of Nephrology, University of Messina, Messina, Italy. [27]Department of Medicine, University of Chieti-Pescara, SS. Annunziata Hospital, Chieti, Italy. [28]S. Andrea Hospital, La Spezia, Italy. [29]University of Bari, Bari, Italy. [30]Department of Nephrology, 1st Faculty of Medicine and General University Hospital, Charles University, Prague, Czech Republic. [31]Second Division of Nephrology, ASST-Spedali Civili di Brescia Presidio di Montichiari, Brescia, Italy. [32]Department of Obstetrics and Gynecology, ASST Spedali Civili di Brescia, Brescia, Italy. [33]Unit of Nephrology and Dialysis ASL TO4, Cirié, Turin, Italy. [34]San Raffaele Hospital, Milan, Italy. [35]Cannizzaro Hospital, Catania, Italy. [36]Sandro Pertini Hospital, Rome, Italy. [37]San Giovanni Bosco Hospital (ERK-net Member) and University of Turin, Turin, Italy. [38]Renal Unit, Department of Medicine,

University of Verona, Verona, Italy. [39]Division of Nephrology, Dialysis, Transplantation, IRCCS Giannina Gaslini, Genoa, Italy. [40]Department of Microbiology and Immunology, Laboratory of Molecular Immunology, Rega Institute, KU, Leuven, Belgium. [41]Department of Nephrology, University Hospitals Leuven, Leuven, Belgium. [42]Department of Family Medicine, Seoul National University College of Medicine and Seoul National University Hospital, Seoul, Korea. [43]Institute on Aging, Seoul National University College of Medicine, Seoul, Korea. [44]Biomedical Sciences, Seoul National University College of Medicine, Seoul, Korea. [45]Kidney Research Institute, Seoul National University College of Medicine, Seoul, Korea. [46]Internal Medicine, Seoul National University College of Medicine, Seoul, Korea. [47]Department of Nephrology, Juntendo University Faculty of Medicine, Tokyo, Japan. [48]Department of Medical Sciences, University of Torino and Immunogenetics and Transplant Biology Service, University Hospital "Città della Salute e della Scienza di Torino", Turin, Italy. [49]Department of Nephrology, University of Toronto, Toronto General Hospital, Toronto, ON, Canada. [50]Division of Nephrology and Hypertension, Mayo Clinic, Rochester, MN, USA. [51]Department of Nephrology and Dialysis, G. Brotzu Hospital, Cagliari, Italy. [52]Manchester Institute of Nephrology and Transplantation, Manchester University Hospitals NHS Trust, Manchester, UK. [53]Department of Nephrology, Division of Medicine, University College London, London, UK. [54]Renal Unit, Royal London Hospital, Barts Health, Whitechapel, London, UK. [55]The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [56]The Genetics of Obesity and Related Metabolic Traits Program, The Icahn School of Medicine at Mount Sinai, New York, NY, USA. [57]The Mindich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [58]Department of Genetics and Genomic Sciences, Mount Sinai Health System, New York, NY, USA. [59]Center for Population Genomic Health, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [60]Institute of Genetic Epidemiology, Dep. of Biometry, Epidemiology, and Medical Bioinformatics, Faculty of Medicine and Medical Center – University of Freiburg, Freiburg, Germany. [61]Department of Nephrology and Medical Intensive Care, Charité – Universitätsmedizin Berlin, Berlin, Germany. [62]Department of Nephrology and Hypertension, Friedrich-Alexander-Universität, Erlangen, Germany. [63]Department of Nephrology, Radboud University Medical Center, Nijmegen, The Netherlands. [64]Department of Human Genetics, Radboud University Medical Centre, Radboud Institute for Health Sciences, Nijmegen, The Netherlands. [65]Department of Epidemiology, Biostatistics & HTA, Radboud University Medical Center, Nijmegen, The Netherlands. [66]Department of Anatomic Pathology, University of Washington, Seattle, USA. [67]Department of Medicine, Renal Section, Boston University School of Medicine and Boston Medical Center, Boston, MA, USA. [68]Institut National de la Santé et de la Recherche Médicale, Centre for Research in Epidemiology and Population Health, Villejuif, France. [69]University Paris-Sud, Villejuif, France. [70]Sorbonne Université, Pierre and Marie Curie University Paris 06, Paris, France. [71]Institut National de la Santé et de la Recherche Médicale, Unité Mixte de Recherche (UMR) 1155, Paris, France. [72]Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, NY, USA. [73]III Department of Medicine, University Medical Center Hamburg-Eppendorf, Hamburg, Germany. [74]Faculty of Biology, Medicine, Health, University of Manchester, Manchester, UK. [75]University of Brescia, Brescia, Italy. [76]Peking-Tsinghua Center for Life Sciences, Beijing, China. [77]These authors contributed equally: Jingyuan Xie, Lili Liu, Nikol Mladkova, Robert Kleta, Nan Chen, Krzysztof Kiryluk. ✉email: nephroxie@163.com; r.kleta@ucl.ac.uk; kk473@columbia.edu

Genetic Identification of Two Novel Loci Associated with Steroid-Sensitive Nephrotic Syndrome

Article *in* Journal of the American Society of Nephrology · July 2019

DOI: 10.1681/ASN.2018101054

CITATIONS
22

READS
112

35 authors, including:

Stephanie Dufek
University College London
36 PUBLICATIONS **338** CITATIONS
SEE PROFILE

Monika Mozere
University of London
15 PUBLICATIONS **224** CITATIONS
SEE PROFILE

Aoife Waters
University College London
27 PUBLICATIONS **564** CITATIONS
SEE PROFILE

Hazel Webb
Great Ormond Street Hospital for Children NHS Foundation Trust
22 PUBLICATIONS **219** CITATIONS
SEE PROFILE

Genetic Identification of Two Novel Loci Associated with Steroid-Sensitive Nephrotic Syndrome

Stephanie Dufek,[1] Chris Cheshire,[1] Adam P. Levine,[1] Richard S. Trompeter,[1] Naomi Issler,[1] Matthew Stubbs,[1] Monika Mozere,[1] Sanjana Gupta,[1] Enriko Klootwijk,[1] Vaksha Patel,[2] Daljit Hothi,[2] Aoife Waters,[2] Hazel Webb,[2] Kjell Tullus,[2] Lucy Jenkins,[2] Lighta Godinho,[2] Elena Levtchenko,[3] Jack Wetzels,[4] Nine Knoers,[5] Nynke Teeninga,[6] Jeroen Nauta,[6] Mohamed Shalaby,[7] Sherif Eldesoky,[7] Jameela A. Kari,[7] Shenal Thalgahagoda,[8] Randula Ranawaka,[8] Asiri Abeyagunawardena,[8] Adebowale Adeyemo,[9] Mark Kristiansen,[10] Rasheed Gbadegesin,[11] Nicholas J. Webb,[12,13] Daniel P. Gale,[1] Horia C. Stanescu,[1] Robert Kleta,[1] and Detlef Bockenhauer[1]

Due to the number of contributing authors, the affiliations are listed at the end of this article.

ABSTRACT

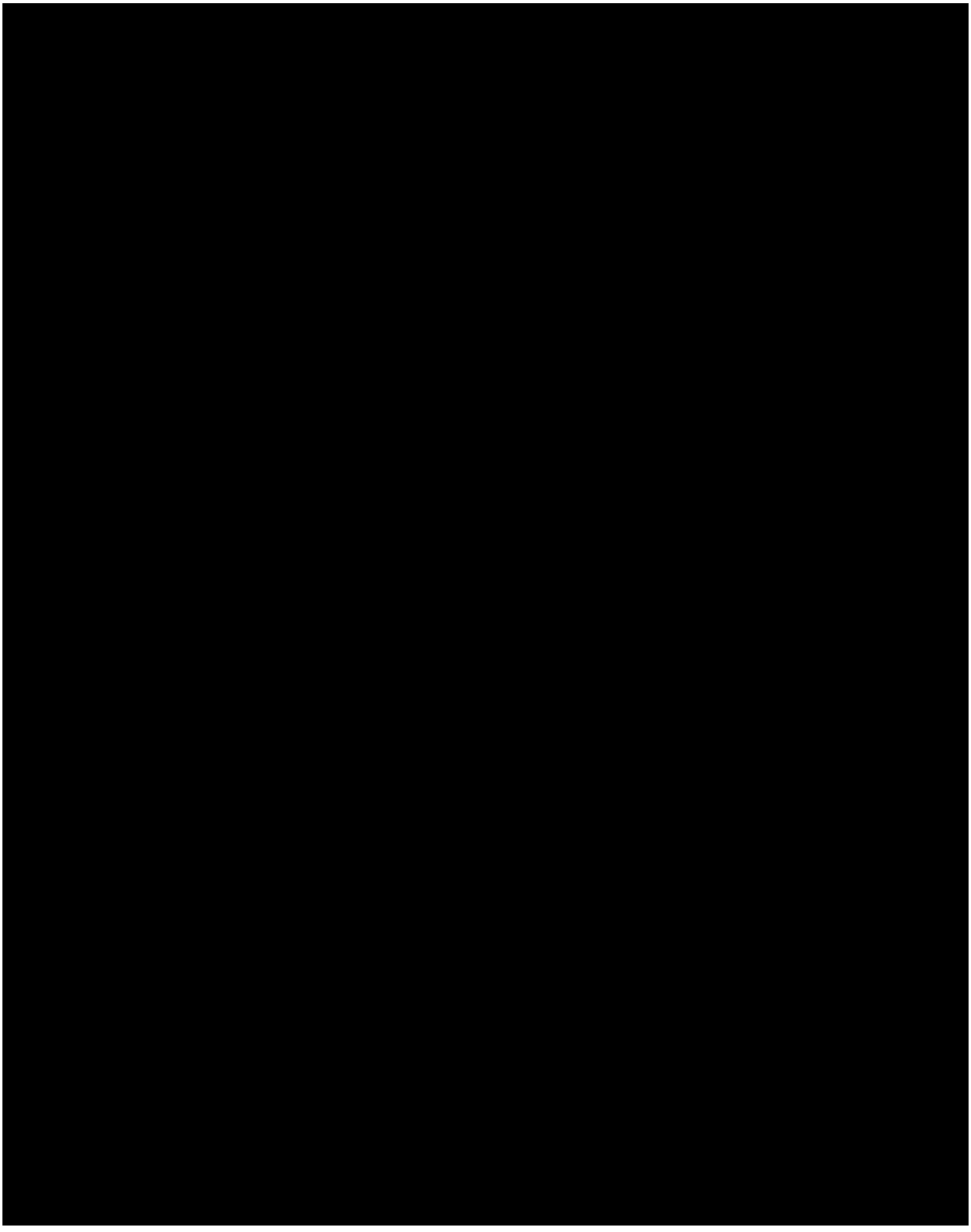
Background Steroid-sensitive nephrotic syndrome (SSNS), the most common form of nephrotic syndrome in childhood, is considered an autoimmune disease with an established classic HLA association. However, the precise etiology of the disease is unclear. In other autoimmune diseases, the identification of loci outside the classic HLA region by genome-wide association studies (GWAS) has provided critical insights into disease pathogenesis. Previously conducted GWAS of SSNS have not identified non-HLA loci achieving genome-wide significance.

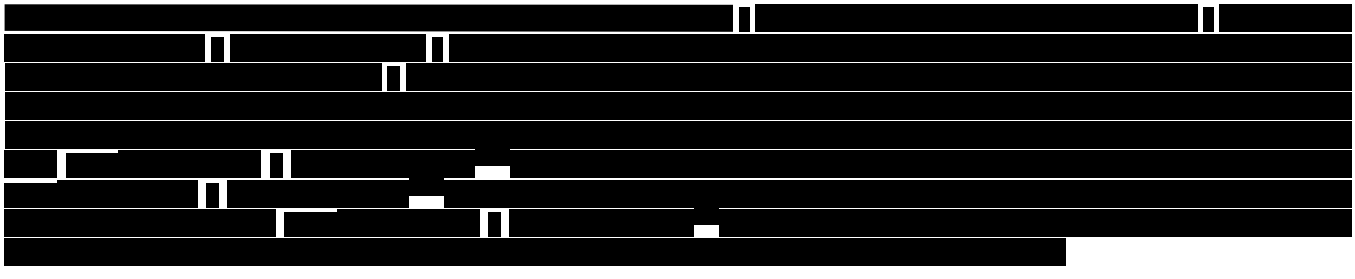
Methods In an attempt to identify additional loci associated with SSNS, we conducted a GWAS of a large cohort of European ancestry comprising 422 ethnically homogeneous pediatric patients and 5642 ethnically matched controls.

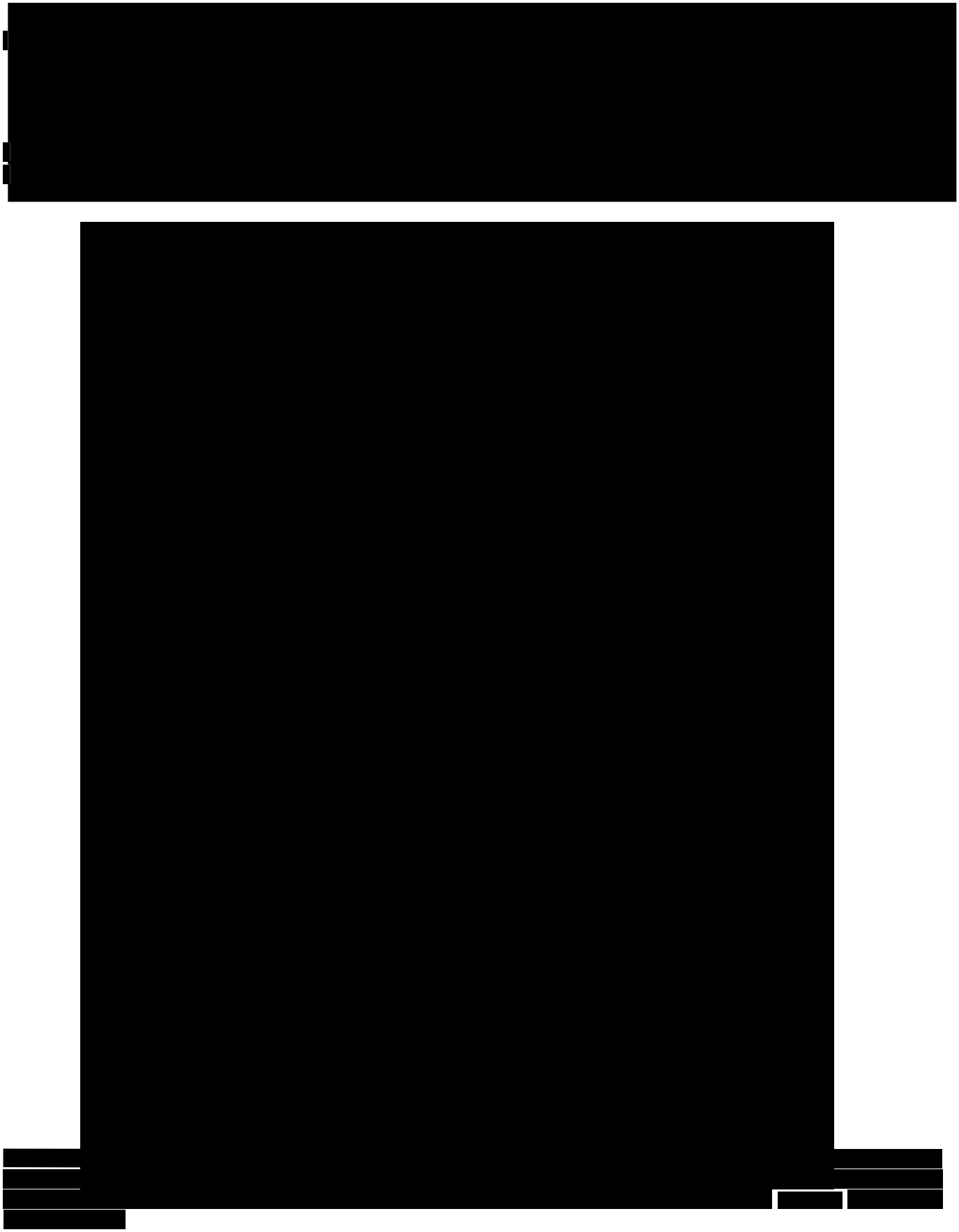
Results The GWAS found three loci that achieved genome-wide significance, which explain approximately 14% of the genetic risk for SSNS. It confirmed the previously reported association with the HLA-DR/DQ region (lead single-nucleotide polymorphism [SNP] rs9273542, $P=1.59\times10^{-43}$; odds ratio [OR], 3.39; 95% confidence interval [95% CI], 2.86 to 4.03) and identified two additional loci outside the HLA region on chromosomes 4q13.3 and 6q22.1. The latter contains the calcium homeostasis modulator family member 6 gene *CALHM6* (previously called *FAM26F*). *CALHM6* is implicated in immune response modulation; the lead SNP (rs2637678, $P=1.27\times10^{-17}$; OR, 0.51; 95% CI, 0.44 to 0.60) exhibits strong expression quantitative trait loci effects, the risk allele being associated with lower lymphocytic expression of *CALHM6*.

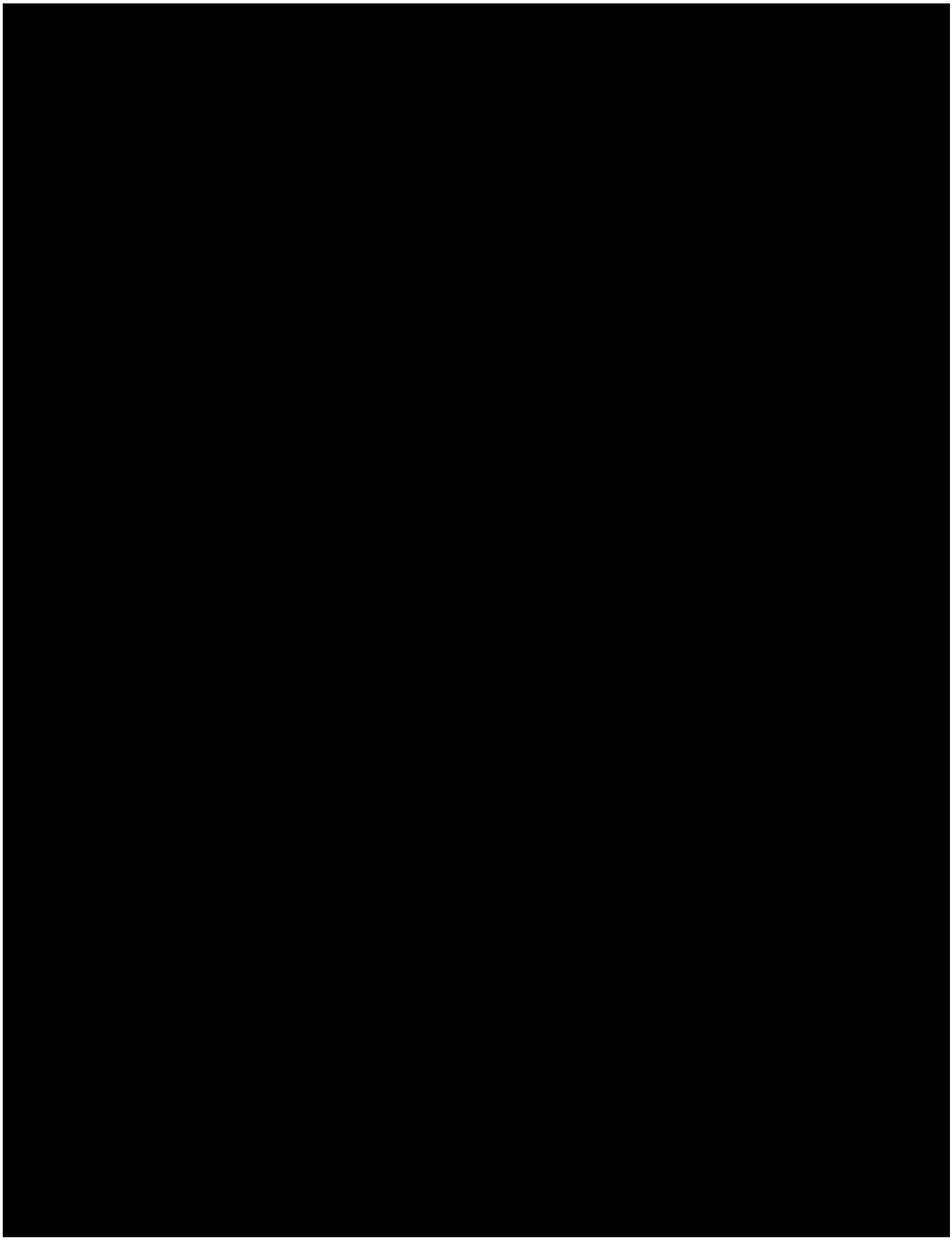
Conclusions Because *CALHM6* is implicated in regulating the immune response to infection, this may provide an explanation for the typical triggering of SSNS onset by infections. Our results suggest that a genetically conferred risk of immune dysregulation may be a key component in the pathogenesis of SSNS.

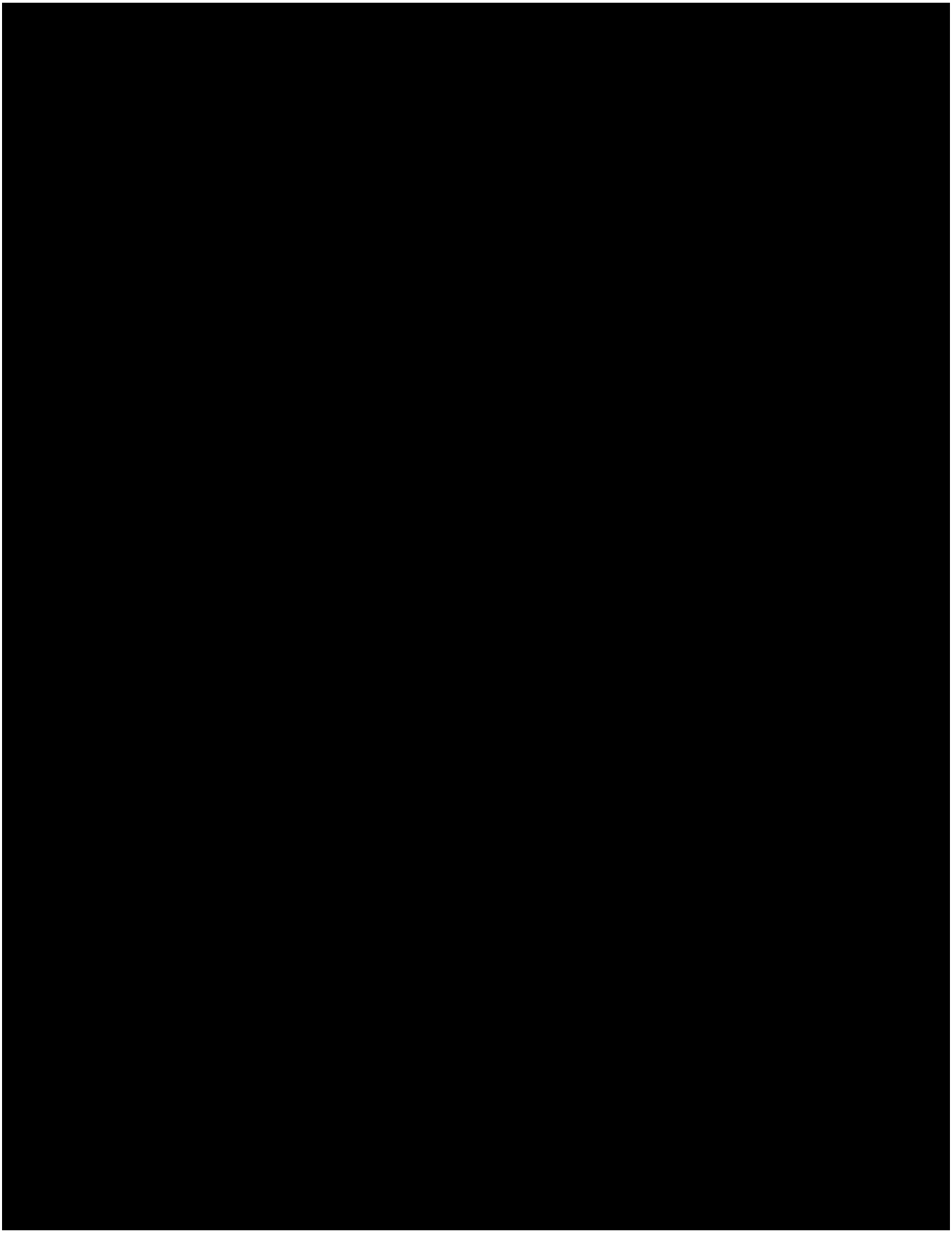
JASN 30: ●●●–●●●, 2019. doi: https://doi.org/10.1681/ASN.2018101054

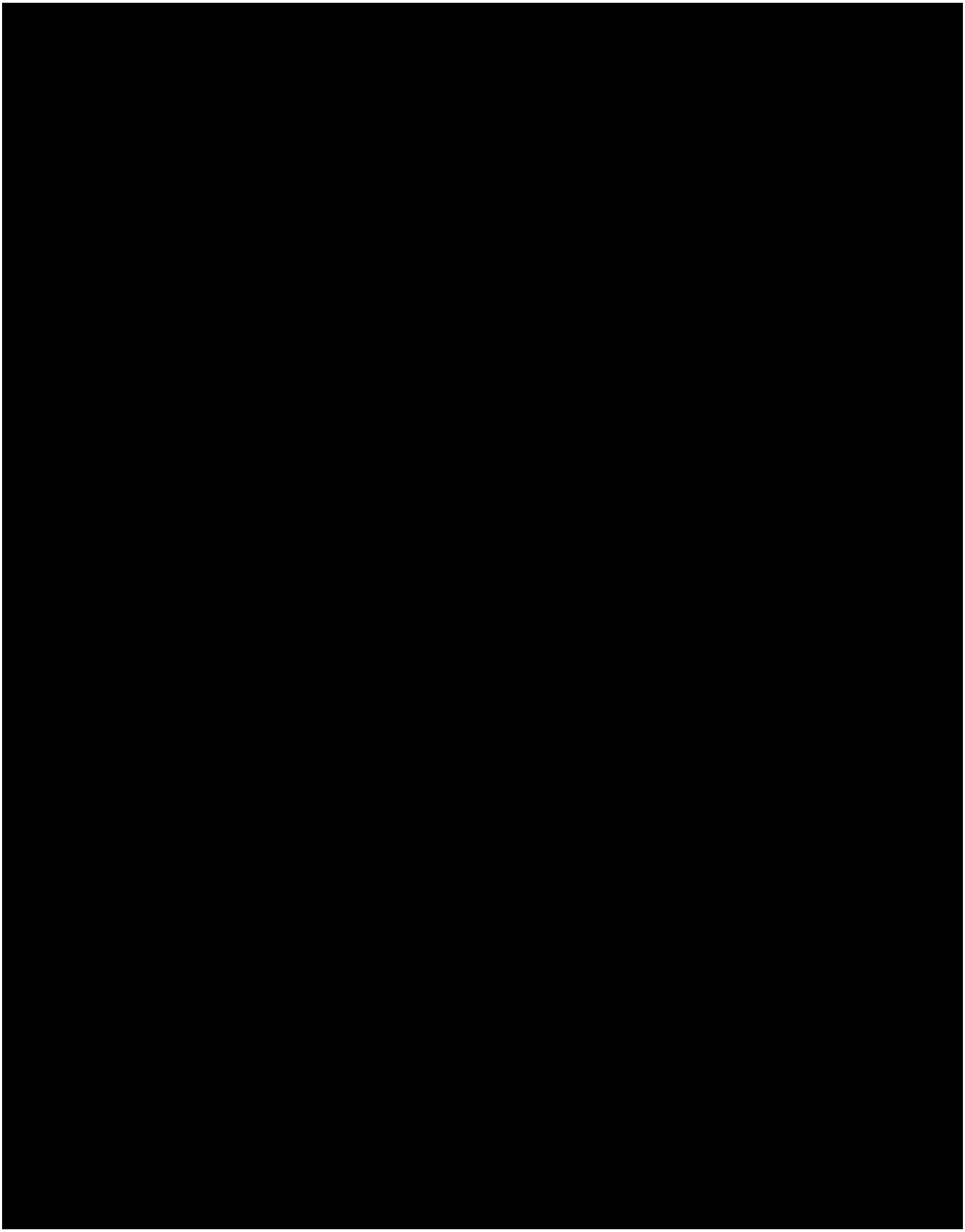


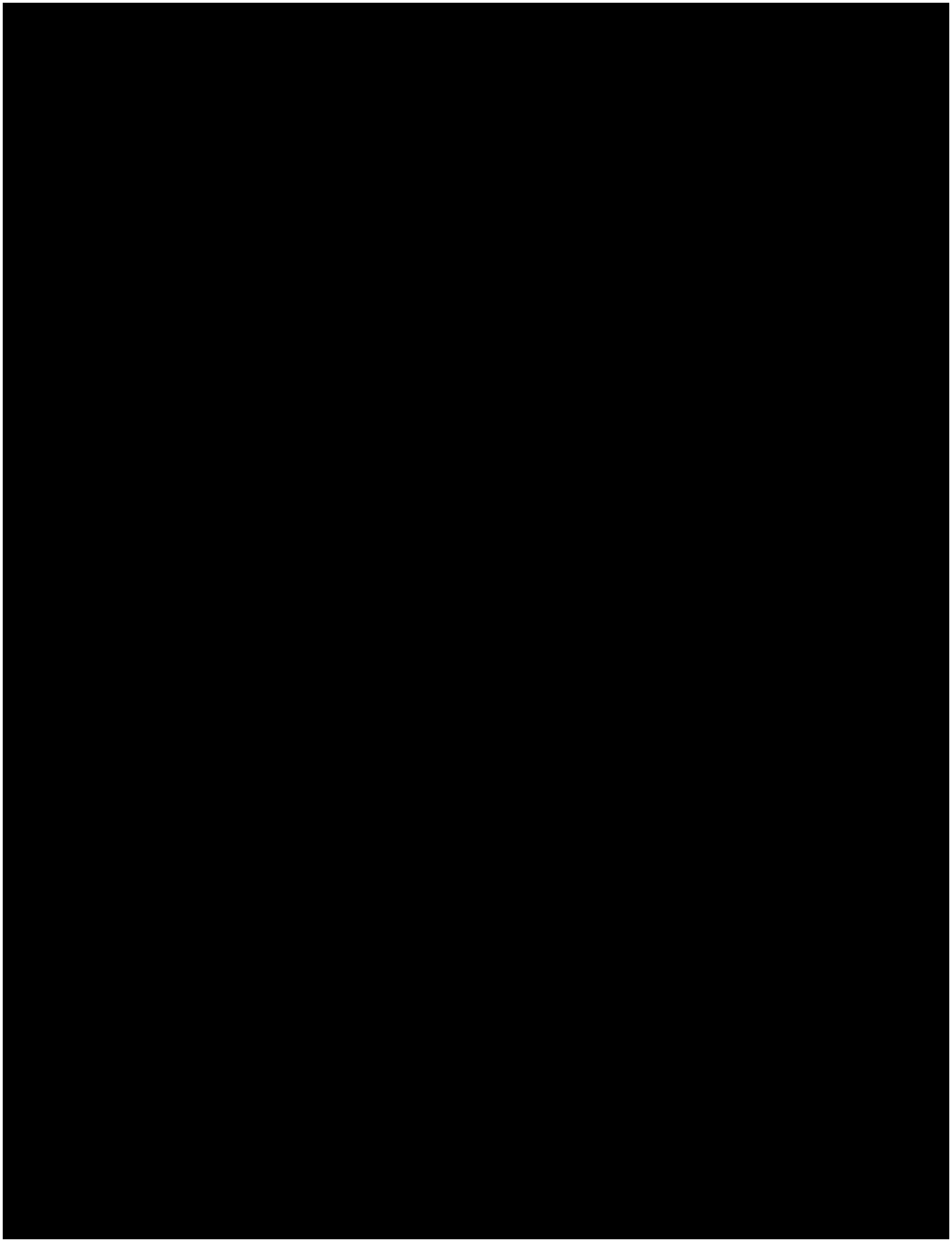












AFFILIATIONS

[1]Department of Renal Medicine and [10]University College London Genomics, Institute of Child Health, University College London, London, United Kingdom; [2]Great Ormond Street Hospital, London, United Kingdom; [3]University Hospitals Leuven and University of Leuven, Leuven, Belgium; [4]Nephrology, Radboud University Medical Center, Nijmegen, The Netherlands; [5]Department of Genetics, UMC Groningen, Groningen, The Netherlands; [6]Department of Pediatric Nephrology, Erasmus University Medical Centre–Sophia Children's Hospital, Rotterdam, The Netherlands; [7]Pediatric Nephrology Center of Excellence, King Abdulaziz University, Jeddah, Kingdom of Saudi Arabia; [8]Department of Paediatrics, University of Peradeniya, Peradeniya, Sri Lanka; [9]NHGRI, National Institutes of Health, Bethesda, Maryland; [11]Department of Pediatrics, Duke University School of Medicine, Durham, North Carolina; and [12]Department of Paediatric Nephrology and [13]NIHR Manchester Clinical Research Facility, Manchester Academic Health Science Centre, Royal Manchester Children's Hospital, Manchester, United Kingdom

Mathew et al. BMC Nephrology (2021) 22:324
https://doi.org/10.1186/s12882-021-02511-x

BMC Nephrology

CASE REPORT Open Access

A case report of breast cancer and membranous nephropathy with positive anti phospholipase A2 receptor antibodies

David Mathew*, Sanjana Gupta and Neil Ashman

Abstract

Background: Testing for antibodies against podocyte phospholipase A2 receptor-1 (PLA2R) allows clinicians to accurately identify primary membranous nephropathy (MN). Secondary MN is associated with a spectrum of pathology including solid organ malignancy. PLA2R positivity in these patients occurs, although no case of PLA2R-positive MN has been definitively linked to cancer.

Case presentation: We describe a case of biopsy-proven PLA2R-positive MN, in whom invasive ductal carcinoma of the breast was discovered. The patient underwent surgery and adjuvant chemotherapy (including cyclophosphamide) and went into a sustained complete remission of her nephrotic syndrome.

Discussion and conclusions: Case series have reported PLA2R positivity in patients with solid organ malignancy associated MN. Our case is unusual as it is a breast malignancy, and the patients nephrotic syndrome and anti-PLA2Rab titres improved with treatment of the cancer. Here we report, to the best of our knowledge, the first case of oestrogen receptor-2 positive breast cancer associated with PLA2R positive MN in a young lady that was treated successfully by treating the malignancy.

Keywords: Cyclophosphamide, Malignancy, Membranous, Nephrotic, PLA2R, Primary, Remission

Background

Antibodies against podocyte phospholipase A2 receptor-1 (PLA2R [1]) were discovered in 2009. Testing for PLA2R antibody allows clinicians to quickly and accurately (specificity approaching 100% [2]) identify primary membranous nephropathy (MN). Secondary MN is associated with a spectrum of pathology including solid organ malignancy. PLA2R positivity in these patients occurs, although no case of PLA2R-positive MN has been definitively linked to cancer [3]. We describe a case of biopsy-proven PLA2R-positive MN, in whom invasive ductal carcinoma of the breast was discovered. The patient underwent surgery and adjuvant chemotherapy (including cyclophosphamide) and went into a sustained complete remission of her nephrotic syndrome.

Case report

A 42 year old Black British woman with no previous medical history of note presented with the nephrotic syndrome (albumin 28 g/L, urine protein creatinine ratio (uPCR) 650 mg/mmol and cholesterol 11.3 mmol/L). Excretory renal function was preserved with estimated glomerular filtration rate (eGFR) > 60 mL/min/1.73m^2.

She described 2 months of leg swelling with no other associated symptoms; physical examination identified ankle oedema and hypertension with a blood pressure of 152/82 mmHg.

Further laboratory testing to investigate her nephrotic syndrome was as follows: Hepatitis B Surface Antigen negative, Hepatitis B Core Antibody negative, Hepatitis C Antibody

* Correspondence: David.mathew2@nhs.net
Department of Nephrology, Royal London Hospital, Whitechapel Road, London E1 1FR, UK

and HIV serology negative. Anti-Nuclear Antibody negative, Extractable Nuclear Antigen negative, Double stranded DNA negative and Rheumatoid Factor undetectable. Immunoglobulin A 2.33 g/L, Immunoglobulin G 7.8 g/L, IgG Subclass 4 0.349 g/L, Immunoglobulin M 0.96 g/L, C3 1.46 g/L, C4 0.47 g/L. No light chains detected on serum or urine protein electrophoresis.

An anti-PLA2R antibody titre was measured at 178kunits/L by ELISA.

Renal biopsy demonstrated characteristic capillary loop thickening, spike formation on silver stain and positive immunohistochemistry for anti-PLA2Rab with polytypic IgG4. A diagnosis of primary MN was made.

Her blood pressure and volume overload were controlled on irbesartan and furosemide. Anticoagulation was declined by the patient even when her albumin dropped to < 25 g/L. The expected hypercholesterolaemia was managed with atorvastatin.

Despite maximal non-immunosuppressive antiproteinuric treatment the patient's nephrosis persisted, and worsened. Her serum albumin fell to 18 g/L, uPCR increased to 950 mg/mmol and anti-PLA2Rab rose on serial testing to 448kunits/L. Eleven months after her initial presentation, in this context, it was agreed with her to treat with immunosuppression. Initiation of this regime was delayed at the patients request. Two months after this decision had been made and, prior to the commencement of any immunosuppression therapy, the patient was diagnosed with multifocal grade 2 invasive ductal carcinoma of the right breast. This was estrogen receptor positive and human epidermal growth factor negative and staging revealed no metastatic disease (pT2 pN1 M0).

She underwent curative treatment with a right mastectomy and axillary lymph node clearance followed by chemotherapy and chest wall radiotherapy. Postoperatively and prior to adjuvant chemotherapy with intravenous cyclophosphamide and doxorubicin she remained nephrotic. She then completed 6 cycles of chemotherapy and received a total cyclophosphamide dose of 6.4 g with doxorubicin 0.64 g.

Clinical improvement of MN timed to successful treatment of the underlying malignancy. After completion of chemotherapy her serum albumin had increased to 34 g/L, the uPCR had improved to 512 mg/mmol (peak 1400 mg/mmol) and the anti-PLA2Rab titre fell to 4kunits/L (peak titre 674kunits/L). Now, 18 months after completing therapy, her anti-PLA2Rab titre is < 2, with a normal serum albumin and a reducing urine PCR of 344 mg/mmol. She is now in a sustained partial remission from her MN.

Discussion and conclusions

Case series have reported PLA2R positivity in patients with solid organ malignancy associated MN. In one [3], only 3 of 10 patients were positive both for serum anti-PLA2RAb and histological IgG4. These patients had stomach, lung and larynx malignancies. Our case is unusual as it is a breast malignancy, and her nephrotic syndrome and anti-PLA2Rab titres improved with treatment of the cancer. Additionally, our patient is young whereas the mean age of malignancy associated MN is 66.

The cyclophosphamide dose used to treat the breast cancer was a lower dose than that used to successfully treat primary MN; however the contribution of this treatment to the resolution of her nephrosis cannot be completely excluded, indeed there are reports of partial remission of MN with Cyclophosphamide doses of less than 3 g [4]. Although less likely given her high PLA2R titre, a spontaneous remission of primary MN is also possible independent of the malignancy.

Here we report, to the best of our knowledge, the first case of oestrogen receptor-2 positive breast cancer associated with PLA2R positive MN in a young lady that was treated successfully by treating the malignancy. We caution clinicians that the exclusive use anti-PLA2Rab in determining a diagnosis of primary MN may not be appropriate.

Case series have demonstrated an association between THSD7A and malignancy in MN [5] and the advent of laser capture microdissection and mass spectrometry has led to the identification of NELL1 as a putative biomarker for malignancy associated MN in PLA2R negative patients [6]. It is likely that further targets will be identified in the field of MN in the coming years which will further elucidate the association between this disease and malignancy.

Abbreviations

eGFR: Estimated Glomerular Filtration Rate; PLA2R: Phospholipase A2 Receptor-1; MN: Membranous Nephropathy; uPCR: urine Protein Creatinine Ratio; NELL1: Nerve Epidermal Growth Factor Like 1; THSD7A: Thrombospondin type-1 domain containing 7A

Acknowledgements

N/A.

Authors' contributions

DM, SG and NA all contributed to literature search and writing up of case report. All authors read and approved the final manuscript. NA was the responsible clinician for this patients care.

Funding

No funding was obtained for this study.

Availability of data and materials

Data referred to from previously published work is referenced in the body of the text.

Declarations

Ethics approval and consent to participate

No ethics approval was required for this case presentation.

Consent for publication
Written consent was obtained was obtained for the publication of this case report from the patient.

Competing interests
The authors declare no conflict of interests.

Received: 15 June 2021 Accepted: 26 August 2021
Published online: 30 September 2021

References
1. Beck LH. Jr1, Bonegio RG, Lambeau G, Beck DM, Powell DW, Cummins TD, et al. M-type phospholipase A2 receptor as target antigen in idiopathic membranous nephropathy. N Engl J Med. 2009;361(1):11–21. https://doi.org/10.1056/NEJMoa0810457.
2. Du Y, Li J, He F, et al. The diagnosis accuracy of PLA2R-AB in the diagnosis of idiopathic membranous nephropathy: a meta-analysis. PLoS One. 2014; 9(8):e104936. https://doi.org/10.1371/journal.pone.0104936.
3. Qin W, Beck LH Jr, Zeng C, Chen Z, Li S, Zuo K, et al. Anti-phospholipase A2 receptor antibody in membranous nephropathy. J Am Soc Nephrol. 2011; 22(6):1137–43. https://doi.org/10.1681/ASN.2010090967 Epub 2011 May.
4. Luzardo L, Ottati G, Cabrera J, Trujillo H, Garau M, Caorsi, et al. Substitution of oral for intravenous cyclophosphamide in membranous nephropathy. Kidney360. 2020;1(9):943–9.
5. Hoxha E, Wiech T, Stahl PR, Zahner G, Panzer U, Thomas NM, et al. A Mechanism for Cancer-Associated Membranous Nephropathy. N. Engl. J. Med. 2016;374:1995–6.
6. Caza T, Hassen S, Dvanajscak Z, Kuperman M, Edmondson R, Larsen P, et al. NELL1 is a target antigen in malignancy-associated membranous nephropathy. Kidney Int. 2021;99(4):967–76. https://doi.org/10.1016/j.kint.2020.07.039 Epub 2020 Aug 20.

Publisher's Note

Rituximab dosing in Membranous Nephropathy may be suboptimal

Lina Nikolopoulou[1], Sanjana Gupta[2], Caroline Tulley[3], Jack Stuart[1], Ruth Pepper[3], Neil Ashman[2], Megan Griffith[1]

[1]Imperial College Healthcare NHS Trust, [2]Barts Health NHS Trust, [3]Royal Free London NHS Trust

BACKGROUND

Rituximab (RTX) is commonly used for the treatment of Membranous Nephropathy (MN) but maintenance of remission and relapse following treatment is variable. Response to RTX after previous immunosuppression (IS) for MN is not well established.

METHODS

91 patients from 3 centres with **biopsy proven MN** treated **with RTX 2x1 gr ,** 2 weeks apart were included. 80/91 (87%) were anti-PLA2R antibody positive. The group was further analysed in two subgroups depending on whether RTX was administered for new presentation of nephrotic syndrome (NS) or for relapsing/ previously treated MN. Baseline characteristics shown on Table 1.

Group 1: Patients with new presentation MN

n=50 patients Follow up (FU) 15 months (median, range 1-34) **44 (88%)** anti PLA2R +ve

Group 2: Relapsing or disease resistant to previous immunosuppression (IS)

n=41, **FU** 8 months (1-38), **36/41(87%)** anti-PLA2R +ve

29/41 previous tacrolimus (TAC) (18/29 – on TAC at time of RTX)

12/41previous cyclophosphamide (CYC)

RESULTS

Time to Bcell Depletion from 1st dose of RTX:

New MN: 4.5 weeks (median, range 1.5 - 15)

Previous TAC: 5.1 weeks (1-24)

Previous Cyclo: 4.5 weeks (3-37) **p:ns**

Time to Bcell Reconstitution from 1st dose of RTX

New MN : 26 weeks the IS naïve group (7.6 -131 w)

Previous treatment: 34 weeks (15-63) **(p:0.1 ns)** (Figure 1)

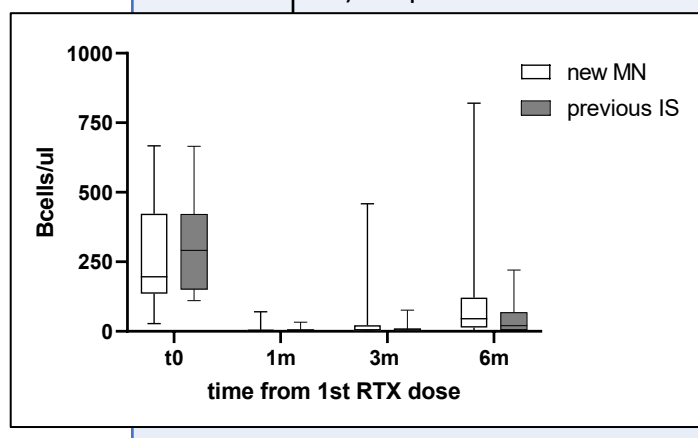


Figure 1. Time to Bcell depletion(Bcell <10cells/ul) following first dose of Rituximab(RTX)

Partial remission (PR): reduction of uPCR >50% from max uPCR<350mg/mmol and >5mg/mmol
Complete remission (CR): uPCR <50mg/mmol and normal serum albumin

At administration of RTX	UPCR mg/mmol median(range)	GFRml/min/1.73m^2 median(range)	anti-PLA2R Ab U/L median (range)	Bcell / uL (median, range)
Total group n=91	899 (155-2363)	60 (13-90)	115(50-1786)	
RTX for New MN n=50	1015(216-2363)*	60(13-90)	172(59-1786)	208 (64-603)
RTX for relapsing/ resistant MN n=41	641(155-1731)* *p:0.002	55(20-90)	71(20-993)	289 (110-665)

Table 1 Baseline proteinuria (urinary protein: creatinine ratio, uPCR) , renal function (glomerular filtration rate ,GFR) and anti-phospholipase A2 receptor (anti-PLA2R) antibody levels of patients treated with Rituximab(RTX) for Membranous Nephropathy at the time of administration of RTX *(p:0.002)

RESULTS

REMISSION

New MN: PR: 26/50 (52%) at 7 months (2-25) from 1st dose RTX

- **5 months (1-24)* from Bcell depletion**
- **15/50 cases: extra dose of RTX at** 30weeks (21-41) from 1st dose
- **CR:4/50 (8%) at 18 months (9-23)** from 1st dose RTX

PREVIOUSLY TREATED : PR 26/41 (63%) (p: 0.3 ns)

Previous TAC: PR: 19/29 (65%) at 3.5 months (1-21)* from 1st RTX

- 2months (1- 20) from Bcell depletion
- 15/18 (83%) cases also on TAC at time of RTX achieved PR after RTX]
- **CR : 6/29 (20%) at 6.5 months (1-17)** from 1st RTX
- 5 m (2.5-16) from B cell depletion

Previous CYC: **PR 7/12 (58%)** at 9m (1-24)* from 1st dose RTX CR: none to date *p >0.5 ns in time to PR

RELAPSE

NEW MN : 4/26(15%) relapsed at a median of 7w (3w -21 m) from reconstitution of Bcells

Previous treatment: TAC:4/ 19 and CYC: 4/7 relapsed median 25w (4-76) – 2 prior to reconstitution of Bcells

CONCLUSIONS

Remission of NS after standard dosing of RTX was not achieved in a significant number of patients in this cohort, although some subsequently achieved this following repeated dosing of RTX.

Time to PR was often long and CR rare.

Time to reconstitution of B cells is variable.

More frequent/increased doses of RTX may be necessary to ensure adequate therapy.

Combination of RTX with adjuvant IS may be of value especially for those with severe disease .

The London Membranous Network

Authors: Sanjana Gupta[1,3], Aikaterini Nikolopoulou[2], John Connolly[3], Thomas Oates[1], Ruth J Pepper[3]*, Megan Griffith[2]*, Neil Ashman[1]*
* equal contributors

[1] Barts Health
[2] Imperial College Healthcare
[3] Royal Free London

Introduction

The London Membranous Network is a North London multi-centre collaborative consortium established to develop a clinical & research infrastructure for membranous nephropathy (MN). Established between three large teaching hospitals, Barts Health, Imperial College Healthcare and the Royal Free London, our three centres support a diverse population of over 6 million people.

Our collaborative will observe our cohort longitudinally, characterising and studying prevalent and incident (40 – 60 per annum) idiopathic MN patients. We are collecting detailed demographic, clinical, histopathological, biochemical, immunological and genetic data.

Methods

Informed patients are enrolled with either biopsy-proven MN, or serological (positive anti-phospholipase A2 receptor antibodies [aPLA2Rab] with nephrotic syndrome) autoimmune MN in a minority of cases where biopsy is not possible.

All investigations are part of standard current clinical care. Where not, ethical approval is in place for collection of DNA, histology, urine and serum for agreed studies. Quality of life and impact of disease can be assessed through validated questionnaires. Follow up in the first year will be on a minimum of 3-monthly basis and thereafter 6-monthly.

Results

At present across all three centres we have a total of 383 idiopathic MN patients. 121 adults cared for at Barts, 99 at the Royal Free and 163 at Imperial College. Diagnosis was established ranging between 1978 and January 2019, with follow up ranging from 0 – 40 years. Differences in disease progression, and response to treatment, in different ethnicities are well-described, with our cohort well-positioned to further examine this (ethnicity in London, table 1). We are continuing full data collection in the prevalent cohort, which is limited by these patients being on renal replacement therapy and not under our care in the specialist membranous clinics.

Conclusion

A London-based collaborative caring at scale for a diverse population will offer useful insight into MN, a rare disease. We aim to better understand diagnosis, symptom control, the consequences of disease and its' (non-immunomodulatory and immunomodulatory) treatment, contributing to the UK Rare Diseases Registry (RaDaR) and global studies in partnership. We aim to measure aPLA2Rab in transplant recipients and those on the transplant waiting list so that we can

determine how the aPLA2Rab status affects renal outcomes. We will seek to support and inform our patients, understand patient-reported outcomes, determine disease and progression risk predictors, and explore best strategies to prevent and manage complications such as thromboembolism, dyslipidaemia, oedema, infections and death (table 2).

	Barts Health	**Imperial College Healthcare**	**Royal Free London**
Total number of patients	121	163	99
Gender Male / Female, n (%)	80 (66) / 41 (34)	108 (66) / 55 (34)	74 (75) / 25 (25
Patient ethnicity: White / Black African & Caribbean / South Asian / East Asian / Other, %	41 / 26 / 30 / 1.5 / 1.5	48 / 14 / 35 / 1 / 2	52 / 22 / 18 / 4 / 4
Age at diagnosis, mean (range)	48 (13-79)	51 (20-92)	50 (15-83)
Biopsy proven, n (%)	118 (98)	163 (100)	97 (98)
Anti-PLA2R positive / negative / unknown, %	Diagnosis 27 / 6 / 67 Prevalent 36 / 40 / 24	Diagnosis 68 / 30 / 2 Prevalent 68 / 30 / 2	Diagnosis 13 / 7 / 80 Prevalent 23 / 34 / 43
GFR at diagnosis mL/min, median (range)	74 (9 to >90)	72 (13 to >90)	69 (4 to >90)
Albumin at diagnosis g/L, mean (range)	26 (9 to 45)	20 (7 to 40)	26 (14 to 45)
Urinary protein creatinine ratio at diagnosis mg/mmol, median (range)	888 (90 to 2300)	884 (27 to 2250)	808 (94 to 2005)
Immunosuppression / no immunosuppression / unknown, n (%)	92 (76) / 29 (24) / 0	137 (84) / 23 (14) / 3 (1)	51 (51) / 40 (40) / 9 (9)
First line immunosuppressive agent: Cyclophosphamide / Calcineurin inhibitors (CNI) / Anti-proliferatives (AP) / AP & CNI Rituximab / Steroid monotherapy, n (%)	53 (58) / 13 (14) / 14 (15) / 0 / 7 (8) / 5 (5)	2 (1) / 97 (71) / 0 / 21 (16) / 15 (11) / 2 (1)	15 (30) / 24 (48) / 4 (8) / 0 / 2 (4) / 5 (10)

Table 1: Baseline data on prevalent idiopathic MN patients in the London Membranous Network

Establish a collaborative, integrated investigational infrastructure to offer support, treatment and conduct clinical research in MN

Conduct longitudinal observational cohort studies in biopsy proven MN
Establish pilot studies from the clinical data collected by London Membranous Network
Obtain funding for future studies to investigate treatment options for treating disease and complications arising from MN
Ultimately with enhanced knowledge and access to medication improve the care and outcomes for patients with MN

Table 2: The objectives of London Membranous Network

frontiers in Physiology

MINI REVIEW
published: 15 January 2019
doi: 10.3389/fphys.2018.01868

Nephrotic Syndrome: Oedema Formation and Its Treatment With Diuretics

*Sanjana Gupta[1,2], Ruth J. Pepper[1], Neil Ashman[2] and Stephen B. Walsh[1]**

[1] UCL Centre for Nephrology, University College London, London, United Kingdom, [2] Renal Unit, The Royal London Hospital, Bart's Health NHS Trust, London, United Kingdom

Oedema is a defining element of the nephrotic syndrome. Its' management varies considerably between clinicians, with no national or international clinical guidelines, and hence variable outcomes. Oedema may have serious sequelae such as immobility, skin breakdown and local or systemic infection. Treatment of nephrotic oedema is often of limited efficacy, with frequent side-effects and interactions with other pharmacotherapy. Here, we describe the current paradigms of oedema in nephrosis, including insights into emerging mechanisms such as the role of the abnormal activation of the epithelial sodium channel in the collecting duct. We then discuss the physiological basis for traditional and novel therapies for the treatment of nephrotic oedema. Despite being the cardinal symptom of nephrosis, few clinical studies guide clinicians to the rational use of therapy. This is reflected in the scarcity of publications in this field; it is time to undertake new clinical trials to direct clinical practice.

Keywords: nephrotic syndrome, diuretics, oedema, amiloride, epithelial sodium channel

OPEN ACCESS

Edited by:
Marcelo D. Carattino, University of Pittsburgh, United States

Reviewed by:
Tengis Pavlov, Henry Ford Health System, United States
James A. McCormick, Oregon Health & Science University, United States

***Correspondence:**
Stephen B. Walsh
stephen.walsh@ucl.ac.uk

Specialty section:
This article was submitted to Renal and Epithelial Physiology, a section of the journal Frontiers in Physiology

Received: *21 August 2018*
Accepted: *11 December 2018*
Published: *15 January 2019*

Citation:
Gupta S, Pepper RJ, Ashman N and Walsh SB (2019) Nephrotic Syndrome: Oedema Formation and Its Treatment With Diuretics. Front. Physiol. 9:1868. doi: 10.3389/fphys.2018.01868

INTRODUCTION

Interstitial oedema is present in all individuals with nephrotic syndrome and can be profound, accounting for as much as an additional 30% to an individual's total body weight (Doucet et al., 2007). Oedema is one of the four defining features of the nephrotic syndrome and is the symptom most commonly requiring intervention (Crew et al., 2004). The other three defining features of nephrotic syndrome are hypoalbuminemia, hyperlipidemia, and proteinuria.

The prevalence of nephrotic syndrome in a small historical retrospective study from 1996 in the United States reported that 60% of all renal biopsies were performed to establish a diagnosis after this presentation (Korbet et al., 1996). It is plausible that this number has not significantly changed. The incidence of primary glomerulonephritides causing the nephrotic syndrome between 1980 to 2010 (limited to the three most common histological findings: minimal change disease, membranous nephropathy and focal segmental glomerular sclerosis) in adults was 2.6 per 100,000 per year (McGrogan et al., 2011).

The burden of symptomatic disease is high, mainly with oedema. Parents of oedematous children report increased anxiety and multiple emergency hospital visits (Beanlands et al., 2017). Peripheral oedema may be uncomfortable, leading to functional restraint, restricted leg movement, and impaired mobility (Doucet et al., 2007). Oedema may also cause increased skin tension, resulting in blistering, skin breakage and exudate extrusion, offering an encouraging environment for bacterial infection. The risk of infection is exacerbated by 'nephrotic immunodeficiency'

caused by the urinary loss of immunoglobulins (Ogi et al., 1994) and T-cell transformation dysfunction (Fodor et al., 1982). Further, asymptomatic pulmonary congestion can be present in active nephrosis (Marino et al., 2016) and there is a strong association of cardiovascular risk in an overloaded chronic kidney disease cohort (Hung et al., 2014).

The two main contributors to oedema are the urinary loss of albumin and excessive sodium. Hence, diuretics are used due to the known tubular effects on sodium and water reabsorption. Before we can target and tailor treatment, we first need to understand the underlying physiology of oedema formation.

OEDEMA FORMATION

There are two paradigms of oedema formation in nephrosis: the so-called under-fill and over-fill models; it is thought that these can be present in the same individual at different times over the course of their disease (Perico and Remuzzi, 1993; Humphreys, 1994). Both result in sodium and water retention and increased interstitial fluid volume presenting as oedema. A detailed review of the pathophysiology of oedema formation is outside the scope of this article, and we direct readers to recent reviews (Siddall and Radhakrishnan, 2012; Cadnapaphornchai et al., 2014; Ray et al., 2015; Ellis, 2016).

Under-Fill Theory

Hypoalbuminemia reduces the capillary oncotic pressure and the imbalance of Starling's forces leads to interstitial leakage of fluid and decreased circulating volume (Doucet et al., 2007). The under-fill theory proposes that the decreased circulating volume leads to renal hypoperfusion and activation of the renin-angiotensin-aldosterone system (RAAS) (Perico and Remuzzi, 1993). Stimulation causes avid sodium and water reabsorption (Doucet et al., 2007).

There are data from rat models of nephrosis suggesting that increased activity of the proximal tubular cell sodium/proton antiporter NHE3 may contribute to sodium retention, but this has not been extensively reported (Besse-Eschmann et al., 2002; Klisic et al., 2003). Furthermore, distal delivery of sodium is the same in nephrotic and non-nephrotic kidneys (Ichikawa et al., 1983).

The distal collecting tubule, cortical and outer medullary collecting ducts are the major sites of sodium reabsorption under the control of aldosterone (Mernissi and Doucet, 1983). In nephrosis, cortical collecting ducts demonstrate increased sodium retention associated with stimulation of basolateral Na^+,K^+-ATPase and apical epithelial sodium channel (ENaC) (Féraille et al., 1993; Deschênes and Doucet, 2000). The stimulation of Na^+,K^+-ATPase results from transcriptional induction of subunits and targeting newly synthesized pumps to the basolateral membrane where the pumps are able to reabsorb sodium (Deschênes et al., 2001a). Pre-formed ENaC is also targeted to the apical membrane via non-transcriptional mechanisms (Lourdel et al., 2005; de Seigneux et al., 2006; Kim et al., 2006).

ENaC is a target for aldosterone and is discussed in detail below. Activation of beta-1 adrenoceptors stimulates renin secretion, however, there are no published studies implicating this in nephrotic syndrome. Renal denervation and beta-blocker therapy are of proven benefit in treating oedema of heart failure (Byku and Mann, 2017), but propranolol did not induce a diuresis or natriuresis (Bauer, 1983) suggesting an alternative unexplained mechanism instead of renin suppression alone.

While the under-fill theory still has considerable support, a sizable body of evidence supports other paradigms of oedema formation. If the under-fill theory was solely responsible then RAAS blockers or restoration of circulating volume would be curative treatment. Captopril (RAAS blocker) failed to change urinary sodium excretion despite successful inhibition of aldosterone secretion (Brown et al., 1984). A proportion of individuals have elevated renin and aldosterone profiles in keeping with an under-filled vascular space, yet, many have suppressed renin and aldosterone activity (Meltzer et al., 1979). Measuring the plasma volume in nephrotic individuals using radioactive albumin demonstrated only 2% of the cohort had a low plasma volume (Geers et al., 1984). Two independent groups in the 1980s demonstrated that intravenous infusion of atrial extract or synthetic atrial natriuretic peptide (ANP) resulted in less than 50% of individuals having a natriuretic and diuretic response despite volume repletion (Koepke and DiBona, 1987; Perico et al., 1989). Interestingly, a low serum albumin alone does not appear to mediate renal sodium retention, as individuals with congenital an-albuminemia do not develop oedema (Koot et al., 2004).

One challenge may be identifying the under-filled individual. A group of Belgian investigators tried to find a clinical test to actively differentiate between these under- and over-filled cohorts (Keenswijk et al., 2018). In nephrotic children a high urinary potassium to urinary potassium and sodium ratio ($UrK^+/UrK^+ + UrNa^+$), suggesting secondary hyperaldosteronism, can be a useful test to identify under-filled children that may benefit from intravenous therapy (Keenswijk et al., 2018).

Over-Fill Theory

The over-fill theory states that RAAS activation and imbalance in Starling's forces across the capillaries are an insufficient mechanism to produce oedema and instead, changes in the capillary endothelial filtration barrier are responsible (Doucet et al., 2007).

Capillary Permeability

Capillary permeability is important in determining distribution of fluid between vascular and interstitial compartments. Changes in the capillary basement membrane akin to aging exist in nephrotic syndrome, with thickening of the basement membrane, altered protein composition and increased stiffness; the thickening of the basement membrane increases protein permeability (Kottke and Walters, 2016).

Nephrotic patients had higher calf capillary filtration capacity without evidence of capillary hypertension suggesting that the functional capillary surface area available for exchange is

increased in nephrotic syndrome (Lewis et al., 1998). The capillary filtration capacity is increased almost twofold in nephrotic individuals (Ellis, 2016).

Studies in nephrotic rats have demonstrated defective capillary basement membrane permeability by ferritin accumulation (Farquhar and Palade, 1961). Vascular permeability is increased in guinea pigs when injected with human lymphocyte supernatants from nephrotic patients but not healthy controls (Lagrue et al., 1975b). A permeability increasing factor (such as a circulating lymphokine) was proposed to affect the vasculature irrespective of the albumin or sodium (Lagrue et al., 1975b). It was hypothesized that this permeability factor was a protein that activates the kinin system (Lagrue et al., 1975a). Initial interest in bradykinin was supported by elevated bradykinin levels in abdominal transudates from nephrotic patients (Paskhina et al., 1979) but no further evidence has been published.

The role for vascular hyperpermeability was further demonstrated using technetium labeled albumin in nephrotic human patients and healthy controls (Rostoker et al., 2000). Nephrotic patients had high levels of hyperpermeability which was reversible with steroids and bilbao extract (which reduces capillary permeability) and the authors proposed a vascular permeability factor derived from local immune cells (Rostoker et al., 2000). The search for the permeability factor remains elusive although possible targets are being proposed such as cytokine receptor-like factor 1 with specific therapeutic targets (Savin et al., 2017). Vascular endothelial growth factors (VEGF) are regulators of capillary permeability (Bates, 2010). However, circulating VEGF levels are not different in nephrotic and non-nephrotic children, and rat VEGF expression is not altered, nor does administration of VEGF induce nephrosis (Webb et al., 1999).

Nephrotic ascites (peritoneal fluid accumulation) is likely to develop via similar mechanisms as nephrotic interstitial fluid (Udwan et al., 2016). In nephrotic rats there is a change in capillary permeability with an increased water filtration coefficient in both paracellular and transcellular pathways, with a reduction in the coefficient for proteins (Udwan et al., 2016). This is associated with greater expression of aquaporin-1 (AQP1) in the parietal peritoneum (Udwan et al., 2016). Water permeability is reversible by inhibiting NF-κB and N-acetylcysteine (Udwan et al., 2016). When both are inhibited, the volume of ascites is reduced by 60% suggesting that this water filtration coefficient, possibly mediated by AQP1, is important in ascites formation in these nephrotic animals (Udwan et al., 2016). The association with aquaporin dysregulation has also been demonstrated in nephrotic human kidneys. A study of 54 primary nephrotic individuals determined aquaporin expression by immunohistochemistry in renal biopsy tissue and measured urinary aquaporin. In nephrotics, AQP1 expression was significantly reduced in renal tissue whereas aquaporin-2 (AQP2) expression was increased. Urinary AQP2 was higher in nephrotic individuals, matching histological findings (Wang et al., 2015).

Endocrine Effects

Adrenalectomized, aldosterone-deficient rats with puromycin aminonucleoside (PAN)-inducible nephrotic syndrome with ascites have lower levels of apically expressed ENaC in the collecting duct than non-adrenalectomized controls (de Seigneux et al., 2006). This suggests that oedema and nephrotic syndrome can occur in the absence of aldosterone and aldosterone-dependent apical expression of ENaC. Other endocrine and paracrine candidates have been investigated in nephrotic animals to explain aldosterone-independent sodium retention in nephrosis. Inhibition of insulin like growth factor, tumor necrosis factor-alpha, nitric oxide synthase in rat models did not induce natriuresis, implying that these are not involved (Doucet et al., 2007).

Angiotensin II is known to have a direct effect on sodium retention independent of glomerular filtration rate and aldosterone secretion (Johnson and Malvin, 1977; Miura et al., 2014). Microperfusion of angiotensin II in distal nephron segments of rat kidney stimulated transporter-like and channel-like sodium retention (Wang and Giebisch, 1996).

The role of circulating factors on sodium retention are difficult to predict as models of animal unilateral nephrosis demonstrate only alterations in sodium handling in the affected kidney (Doucet et al., 2007). The main determinant of sodium retention is the Na^+,K^+-ATPase pump (discussed above) and the capillary epithelial dysfunction.

Atrial natriuretic peptide stimulates sodium and water excretion from the inner medullary collecting duct thereby opposing the interstitial volume expansion (Baxter et al., 1988; Light et al., 1989). After prolonged exposure to ANP there is an increase in apical expression of aquaporin-2 and the gamma subunit of ENaC (Wang et al., 2006). However, systemic infusion of synthetic ANP or ANP extract in experimental studies in nephrotic individuals had a diminished natriuretic and diuretic response compared to healthy controls (Koepke and DiBona, 1987; Perico et al., 1989) despite the prediction from animal models for the opposite (Wang et al., 2006).

Regardless of elevated serum ANP concentration in nephrotic syndrome, natriuresis is blunted (Perico et al., 1989). This may be due to increased activity of renal sympathetic nerves that override the ANP response and stimulate sodium retention, although denervated kidneys still demonstrate the same effect (Maack, 1980). Alternatively, dysfunctional ANP binding in the kidney may explain the relative renal resistance to ANP. In a murine model of nephrotic syndrome, renal resistance to ANP is reversed by phosphodiesterase inhibitors, implying that this ANP resistance is mediated through enhanced cyclic GMP–phosphodiesterase activity (Valentin et al., 1992).

The *PPIL* gene encoding cyclophilin-like protein is upregulated in the medulla from nephrotic rats compared with healthy control rat medullas (Orisio et al., 1993). The gene product cyclophilin-like protein reduces sodium excretion (Iwai and Inagami, 1990). ANP infusion increases the cyclophilin-like protein mRNA in the renal medulla in nephrotic rats, so this may be a potential mechanism of ANP insensitivity in nephrosis

(Orisio et al., 1993). Down-regulation of the protease corin in kidneys reduces conversion of pro-ANP to active ANP, also contributing to a lack of renal response to ANP (Polzin et al., 2010).

EPITHELIAL SODIUM CHANNEL

The ENaC mediates absorption of sodium in the late distal convoluted tubule, connecting segment and tubule and the collecting duct (Garty and Palmer, 1997). ENaC is activated by aldosterone, anti-diuretic hormone and specific proteases (Garty and Palmer, 1997; Kleyman et al., 2009; Rossier and Stutts, 2009; Passero et al., 2010). The role of peroxisome proliferator activated receptors oedema and ENaC remains unclear and is reviewed in Pavlov et al. (2010). ENaC is composed of three subunits: alpha, beta and gamma (Firsov et al., 1998). Proteases activate ENaC by cleaving the alpha and gamma subunits (Hughey et al., 2003). Dual cleavage of the gamma subunit renders ENaC highly active (Sheng et al., 2006). Further, animal and human nephrotic urine activates ENaC; proteases (e.g., plasmin) in the urine activate and inhibitors of plasmin deactivate ENaC dependent sodium currents in *Xenopus laevis* oocytes (Svenningsen et al., 2009). Furthermore, remission of nephrotic syndrome is associated with a reduction in urinary plasmin levels in patients' urine. Patients' urine was applied to M1 collecting duct cells expressing ENaC and there was a lower ability to activate ENaC-mediated (amiloride sensitive) sodium currents compared to urine from nephrotic patients (Andersen et al., 2013).

Most recently, a cohort study reports urinary plasmin increases the risk of hypertension in type 1 diabetics. Despite this association, the increased risk was dependent on albuminuria and nor were there differences in urinary sodium or potassium excretion. This was attributed to uncontrolled patient sodium intake and urinary plasminogen deactivating ENaC simultaneously (Ray et al., 2018).

Recent studies in mice using pharmacological inhibitors of protease activity to inhibit the gamma subunit cleavage and activation of ENaC have been successful at preventing sodium retention. Urinary protease inhibitors normalized protease activity and reduced sodium retention (Bohnert et al., 2018). Rats with induced nephrotic syndrome had higher levels of urinary plasminogen activator, and amiloride reduced this and sodium retention without altering proteinuria (Stæhr et al., 2015).

Protease inhibition might be an attractive therapeutic option in addition to ENaC antagonism (e.g., with amiloride). Protease inhibition would occur before glomerular filtration, and therefore its effectiveness would not be limited by the glomerular filtration rate, unlike diuretics which act on the tubular epithelium after filtration. Further, while the onset of action of a putative protease inhibitor will be slower than ENaC antagonism, as it depends on the rate of ENaC retrieval to the apical membrane (Gaillard et al., 2010), the effect is likely to be sustained rather than the short action of ENaC antagonism. Avid sodium retention in between periods of action of diuretics has long been observed (Wilcox et al., 1983; Loon et al., 1989), and is an important contributor to diuretic resistance. Protease inhibition may therefore present an intriguing route to ameliorate diuretic resistance in nephrotic syndrome.

In animal models of nephrotic syndrome amiloride successfully abolishes the abnormally high sodium reabsorption from the cortical collecting duct independent of aldosterone activity (Deschênes et al., 2001b, 2003). ENaC inhibitors (amiloride) are not without side effects; they increase serum potassium (Brown et al., 2016) and pose a risk of hyperkalemia so should be used with caution in advanced chronic kidney disease (Wile, 2012) or diabetic patients (Unruh et al., 2017). When used in combination with other medications such as RAAS blockers and diuretics the risk of acute kidney injury is higher (Unruh et al., 2017; Hinrichs et al., 2018; Ray, 2018). There is also an association with pressure ulcers in hospitalized patients on amiloride (Roustit et al., 2016). There are other potential downstream effects of amiloride. Amiloride inhibits urokinase plasminogen activator and reduces plasmin generation (Vassalli and Belin, 1987). Plasmin is pro-fibrotic (Zhang et al., 2007) and therefore amiloride may potentially affect renal fibrosis. Additionally, podocyte anchoring to the glomerular basement membrane is affected by amiloride in experimental animal studies and thereby could influence proteinuria (Zhang et al., 2012; Reiser, 2013; Trimarchi et al., 2014; Warnock, 2015).

Epithelial sodium channel knockout mice exhibited downregulation of the sodium chloride cotransporter (NCC) which was then unable to compensate with the usual enhanced compensatory sodium reabsorption (Perrier et al., 2016). NCC can be phosphorylated and regulated by the serum potassium concentration (Czogalla et al., 2016; Terker et al., 2016). Reduced ENaC activity induces hyperkalemia which may be a more important regulator of NCC than sodium balance (Boscardin et al., 2018). It is possible that amiloride may mediate a degree of NCC inhibition by increasing the serum potassium.

Combination treatments with amiloride have been trialed. Amiloride causes a significant diuresis in mouse models treated with acetazolamide (Patel-Chamberlin et al., 2016). Amiloride and hydrochlorothiazide improved weight loss compared to placebo in a population aged over 65 (Damian et al., 2016). The addition of amiloride to RAAS blockers and hydrochlorothiazide, improved proteinuria reduction by an additional 14% (Morales et al., 2015). Despite the growing evidence for ENaC inhibition, loop diuretics remain the mainstay of treatment.

We have focused on the putative mechanisms of salt and water retention in nephrotic syndrome; however, mechanisms other than renal sodium retention may be important in causing interstitial oedema.

TREATMENT OF OEDEMA

Diuretics

We refer the reader to the review of Wile for details on the history of diuretics (Wile, 2012). All classes of diuretics act within the kidney to reduce renal tubular sodium reabsorption, limiting water reabsorption with resulting diuresis. Diuretics are classed as osmotic diuretics, carbonic anhydrase inhibitors, loop, thiazide or potassium-sparing diuretics (The National

Institute for Health and Care [NICE], 2018). All diuretics (with the exception of mineralocorticoid receptor antagonists) act on the luminal side of the tubular epithelium and need to attain sufficient concentration to have an action there (Wile, 2012).

Loop diuretics (e.g., furosemide and bumetanide) work by inhibiting the Na^+-K^+-$2Cl^-$ cotransporter, NKCC2 on the apical surface in the thick ascending limb (TAL) in the loop of Henle (Brater, 1991). This transporter reabsorbs sodium (Na^+) in to the tubular cell which is reclaimed to the circulation by the Na^+,K^+-ATPase pump in the basolateral membrane (Brater, 1991). The TAL is the site of approximately 25% of total sodium reabsorption in the nephron, so loop diuretics are particularly potent natriuretics (Burg, 1982).

Loop diuretics are highly protein bound and are secreted in to the lumen by organic anion transporters along the proximal tubule to reach their site of action in the TAL (Brater, 1991). Gut oedema may limit the oral absorption of diuretics and hypoalbuminemia decreases the delivery of diuretic to its site of action (Sica, 2003). For this reason higher doses may be required to achieve a successful diuresis, although with a higher likelihood of adverse effects (Crew et al., 2004). Intravenous infusion is another approach that can increase loop diuretic efficacy as therapeutic levels are achieved rapidly due to lower absorption time (Hammarlund et al., 1984).

Furosemide has incomplete and variable bioavailability even in healthy individuals (Waller et al., 1981); probably the worst in its class, despite its widespread popularity. Furosemide has both poor aqueous solubility and low intestinal permeability, and oral furosemide bioavailability varies greatly (by an estimated 20–60%) both between individuals and within individuals (Granero et al., 2010; Nielsen et al., 2016). In the oedematous state bioavailability is further reduced (e.g., to 30%) (Odlind and Beermann, 1980). There is current work using nanoparticles or polymeric microcontainers aiming to stabilize and reduce this effect (Sahu and Das, 2014; Nielsen et al., 2016).

Any diuretics can cause hypovolemia with secondary acute kidney injury (Crew et al., 2004; Oh and Han, 2015). Electrolyte dysregulation is frequent; including hypokalemia, hypomagnesemia, hypocalcemia, hyponatremia, and hyperuricemia. This requires monitoring. Moreover, furosemide can cause hypersensitivity reactions such as skin rashes or acute interstitial nephritis (Oh and Han, 2015). There is a risk of reversible ototoxicity related to the peak serum drug concentration due to the ubiquitous NKCC1 present in the inner ear also being inhibited with loop diuretic use (Wile, 2012). Diuretics can displace warfarin from its protein binding site, increasing the anti-coagulant effect (Oh and Han, 2015). This is seen with furosemide (Oh and Han, 2015) and the active metabolite (canrenone) of spironolactone (Takamura et al., 1997).

Novel Agents

Vasopressin receptor antagonists (e.g., tolvaptan) are not diuretics, but rather aquaretics. These drugs reduce the density of luminal aquaporins to increase urinary water excretion without natriuresis (The National Institute for Health and Care [NICE], 2018). To date there are three case reports on four individuals describing the use of tolvaptan in massive oedema in nephrotic individuals, with treatment described as successful in three out of four cases (Shimizu et al., 2014; Park et al., 2015; Tanaka et al., 2017).

Phase I trials have been completed for a novel particulate-guanylyl-cyclase A receptor activator (trial name ZD100) that promotes natriuresis, inhibits aldosterone and reduced blood pressure with activation of cyclic guanosine monophosphate (Chen et al., 2016). At present this is being investigated in the use of resistant hypertension but if it has natriuretic properties it may be suitable for use in oedema in nephrosis.

Relaxin, an endogenous neurohormone is currently being trialed in heart failure and has completed phase III trials (Wilson et al., 2015). Relaxin induces increased expression of both epithelial and endothelial endothelin B receptor (ET_B) and thereby indirectly stimulates ET_B (Garvin and Sanders, 1991; Danielson et al., 2000; Bogzil et al., 2005; Schneider et al., 2007). ET_B inhibits Na^+,K^+-ATPase and ADH causing both a natriuresis and diuresis (Garvin and Sanders, 1991). The fractional increase in urinary excretion of sodium occurs without any changes in aldosterone or ANP concentrations (Bogzil et al., 2005). In heart failure, treated individuals were found to require lower doses of loop diuretics, had greater weight loss and reduced signs and symptoms of fluid overload including peripheral and pulmonary oedema (Metra et al., 2013; Voors et al., 2014). However, this has had no effect on overall outcomes (Teerlink, 2017; Teerlink et al., 2017), and nor is it known if the endothelin dysregulation in heart failure is present in nephrosis.

There have been two cases from a single center in Japan of synthetic human ANP (carperitide) reducing interstitial oedema preventing the need for hemodialysis and preserving renal function in nephrotic patients (Ueda et al., 2014).

Luteolin is a common phenolic compound known to have anti-inflammatory and anti-allergic effects. Rat studies have recently demonstrated the role of luteolin in natriuresis and diuresis with an additive effect achieved with administration of amiloride and hydrochlorothiazide (Boeing et al., 2017). Luteolin mediated these effects via the muscarinic acetylcholine receptor (Boeing et al., 2017). It will be interesting to see if this has a similar role to play in humans.

Most recently, the role of epicatechin was investigated. Epicatechin is a flavonoid found in food and plant extracts and is classed as a phytochemical (Mariano et al., 2018). Rats treated with epicatechin achieved diuresis and uresis of sodium, potassium and chloride without any effect on the plasma electrolyte function (Mariano et al., 2018). Combination with hydrochlorothiazide further improved the diuretic effect (Mariano et al., 2018).

Clinical Practice

There are no adult guidelines available on managing oedema and volume overload in nephrotic syndrome (Crew et al., 2004; Oh and Han, 2015). The lack of guidelines means

that there is considerable heterogeneity in the treatment of overloaded nephrotic individuals with no clear consensus. Non-pharmacological interventions such as strict dietary sodium restriction (less than 3g per day) are important and have an additive effect to other therapies (Hull and Goldsmith, 2008). Replacing serum albumin with intravenous infusions to improve the efficacy of loop diuretics has been investigated considerably. This is because only albumin-bound loop diuretics are secreted into the lumen of the tubule, discussed above. Hypoalbuminemia (as is present in nephrotic syndrome) reduces the amount of loop diuretics that are delivered to their site of action in the tubular lumen (Kirchner et al., 1990; Fliser et al., 1999). Albumin may help overcome this by enhancing proximal tubular secretion to the tubular lumen and reducing the volume of distribution (Inoue et al., 1987; Chalasani et al., 2001). Early reports were encouraging with massive diuresis when combined with diuretics (27 kg in 14 days) (Davison et al., 1974). Later studies demonstrate conflicting results with better results in animal models and minimal response in humans. There was little or no effect on urinary furosemide or sodium excretion, despite elevated serum albumin levels (Akcicek et al., 1995; Fliser et al., 1999; Chalasani et al., 2001). Nor did inhibiting urinary protein binding have any effect of furosemide efficacy (Agarwal et al., 2000). Different mechanisms may exist in children as there has been a successful response with co-administration of albumin in some instances (Haws and Baum, 1993). Yet, the rate of complications remains higher in pediatric patients with 38% developing hypertension (Dorhout Mees, 1996). We refer the reader to table 1 in the following review for a full synopsis on all clinical trials with albumin and furosemide (Duffy et al., 2015).

A multicenter retrospective study of 60 pediatric units in Italy found no consensus approach to diuretic treatment (Pasini et al., 2015). Despite similar laboratory and clinical data across 231 children, 64% were treated with diuretics and 55% were treated with albumin infusions, highlighting the treatment variability (Pasini et al., 2015). The authors suggest that shared guidelines and implementation were necessary to avoid differences and side effects in pediatric patients (Pasini et al., 2015).

In adults, there is also no consensus on the indication, starting dose, approach to dosage change and monitoring of diuretics; consequently, there are considerable differences in treatment pathways. In general, standard first line treatment is a loop diuretic such as furosemide (Crew et al., 2004). The diuretic would normally be started at a low dose and then sequentially increased until satisfactory weight loss or until the maximum dose has been reached. If at this stage the patient remains symptomatic, a second, synergetic diuretic is commonly added. In the United Kingdom, this would usually be the thiazide-like metolazone (Crew et al., 2004). This is not always successful and a recent case report described resistance to the loop and thiazide type diuretic combination in a oedematous nephrotic patient. The oedema instead responded well to a combination of furosemide and an ENaC inhibitor (triamterene) (Hoorn and Ellison, 2017). Another case report describes a treatment resistant nephrotic patient with no diuretic effect after 5 weeks of high dose furosemide but a profound 7 kg fluid loss with the addition of amiloride (Hinrichs et al., 2018). While these are only two case reports they provide further evidence to support ENaC blockade as effective treatment of nephrotic oedema, although larger studies will be necessary.

There are no trials directly comparing the commonly used different diuretic regimens in nephrotic syndrome despite the underlying scientific basis of causation for oedema in animal models. Furosemide remains first line treatment yet has extremely variable oral bioavailability and the problematic well-described blunted natriuretic effect with prolonged therapy (Bernstein and Ellison, 2011).

More recently, a randomized controlled study in Iran compared different diuretic pre-loading regimens in 20 individuals with refractory nephrosis. The two arms were either pre-loading with acetazolamide and hydrochlorothiazide compared to pre-loading with furosemide and hydrochlorothiazide for 1 week after which both treatment arms received 2 weeks of 40 mg furosemide. Patients with hypokalemia were excluded in view of the potassium wasting properties of these diuretics. The authors concluded that the acetazolamide and hydrochlorothiazide combination achieved a better diuresis. While there was a difference in the mean weight change and urinary volume there was no difference in the urinary sodium and hence natriuresis over the trial period (Fallahzadeh et al., 2017). We discussed above that this was successful in animal models (Patel-Chamberlin et al., 2016). Inhibition of pendrin which is found in the collecting duct is part of a proposed route of sodium reabsorption in the collecting duct with acetazolamide (Zahedi et al., 2013). This may be another mechanism for the synergistic effect with a loop diuretic (Soleimani et al., 2012).

CONCLUSION

Patients and clinicians deserve better and more structured information on how to successfully manage nephrotic oedema. With increased understanding of the underlying pathophysiology of interstitial oedema in nephrotic syndrome we are in a better position to better treat individuals suffering with the complications that severe interstitial oedema brings. Management of oedema in nephrotic syndrome will rely on individualization of therapy with consideration of the clinical evidence. Differences especially will exist in the treatment between pediatric and adult patients and it is likely that there is no one best treatment for everyone. It will be helpful to understand determinants of response to treatment. Based on the evidence we provide in the review we especially advocate clinical trials to investigate the potential benefits of ENaC blockade.

For example, it may be important in certain types of drug-induced oedema, such as that seen with thiazolidinediones,which stimulate ENaC-mediated sodium reabsorption through PPARγ receptors (Guan et al., 2005). We hope that with appropriately designed trials, guidelines can be developed based on robust clinical evidence to achieve improved outcomes.

AUTHOR CONTRIBUTIONS

SG drafted the manuscript with critical revision by RJP, NA, and SBW.

FUNDING

SG has received funding from UCB for her Ph.D.

ACKNOWLEDGMENTS

We would like to acknowledge the support from the UCL Centre for Nephrology to SBW, RJP, and SG. The support from the glomerular disease group at Barts Health for NA and SG. Additionally SG, RJP, NA, and SBW are grateful to their ongoing collaborations with the London Membranous Network.

REFERENCES

Agarwal, R., Gorski, J. C., Sundblad, K., and Brater, D. C. (2000). Urinary protein binding does not affect response to furosemide in patients with nephrotic syndrome. *J. Am. Soc. Nephrol.* 11, 1100–1105.

Akcicek, F., Yalniz, T., Basci, A., Ok, E., and Mees, E. J. D. (1995). Diuretic effect of frusemide in patients with nephrotic syndrome: is it potentiated by intravenous albumin? *BMJ* 310, 162–163. doi: 10.1136/bmj.310.6973.162

Andersen, R. F., Buhl, K. B., Jensen, B. L., Svenningsen, P., Friis, U. G., Jespersen, B., et al. (2013). Remission of nephrotic syndrome diminishes urinary plasmin content and abolishes activation of ENaC. *Pediatr. Nephrol.* 28, 1227–1234. doi: 10.1007/s00467-013-2439-2

Bates, D. O. (2010). Vascular endothelial growth factors and vascular permeability. *Cardiovasc. Res.* 87, 262–271. doi: 10.1093/cvr/cvq105

Bauer, J. H. (1983). Effects of propranolol therapy on renal function and body fluid composition. *Arch. Intern. Med.* 143, 927–931. doi: 10.1001/archinte.1983.00350050085016

Baxter, J. D., Lewicki, J. A., and Gardner, D. G. (1988). Atrial natriuretic peptide. *Nat. Biotechnol.* 6, 529–546. doi: 10.1038/nbt0588-529

Beanlands, H., Maione, M., Poulton, C., Herreshoff, E., Hladunewich, M. A., Hailperin, M., et al. (2017). Learning to live with nephrotic syndrome: experiences of adult patients and parents of children with nephrotic syndrome. *Nephrol. Dial. Transplant.* 32, i98–i105. doi: 10.1093/ndt/gfw344

Bernstein, P. L., and Ellison, D. H. (2011). Diuretics and salt transport along the nephron. *Semin. Nephrol.* 31, 475–482. doi: 10.1016/j.semnephrol.2011.09.002

Besse-Eschmann, V., Klisic, J., Nief, V., Hir, M. L., Kaissling, B., and Ambühl, P. M. (2002). Regulation of the proximal tubular sodium/proton exchanger NHE3 in rats with puromycin aminonucleoside (PAN)-induced nephrotic syndrome. *J. Am. Soc. Nephrol.* 13, 2199–2206. doi: 10.1097/01.ASN.0000028839.52271.DF

Boeing, T., da Silva, L. M., Mariott, M., Andrade, S. F., and de Souza, P. (2017). Diuretic and natriuretic effect of luteolin in normotensive and hypertensive rats: role of muscarinic acetylcholine receptors. *Pharmacol. Rep.* 69, 1121–1124. doi: 10.1016/j.pharep.2017.05.010

Bogzil, A. H., Eardley, R., and Ashton, N. (2005). Relaxin-induced changes in renal sodium excretion in the anesthetized male rat. *Am. J. Physiol. Regul. Integr. Comp. Physiol.* 288, R322–R328. doi: 10.1152/ajpregu.00509.2004

Bohnert, B. N., Menacher, M., Janessa, A., Wörn, M., Schork, A., Daiminger, S., et al. (2018). Aprotinin prevents proteolytic epithelial sodium channel (ENaC) activation and volume retention in nephrotic syndrome. *Kidney Int.* 93, 159–172. doi: 10.1016/j.kint.2017.07.023

Boscardin, E., Perrier, R., Sergi, C., Maillard, M. P., Loffing, J., Loffing-Cueni, D., et al. (2018). Plasma potassium determines NCC abundance in adult kidney-specific γ ENaC knockout. *J. Am. Soc. Nephrol.* 29, 977–990. doi: 10.1681/ASN.2017030345

Brater, D. C. (1991). Clinical pharmacology of loop diuretics. *Drugs* 41, 14–22. doi: 10.2165/00003495-199100413-00004

Brown, E., Markandu, N., Sagnella, G., Jones, B., and MacGregor, G. (1984). Lack of effect of captopril on the sodium retention of the nephrotic syndrome. *Nephron* 37, 43–48. doi: 10.1159/000183206

Brown, M. J., Williams, B., Morant, S. V., Webb, D. J., Caulfield, M. J., Cruickshank, J. K., et al. (2016). Effect of amiloride, or amiloride plus hydrochlorothiazide, versus hydrochlorothiazide on glucose tolerance and blood pressure (PATHWAY-3): a parallel-group, double-blind randomised phase 4 trial. *Lancet Diabetes Endocrinol.* 4, 136–147. doi: 10.1016/S2213-8587(15)00377-0

Burg, M. B. (1982). Thick ascending limb of Henle's loop. *Kidney Int.* 22, 454–464. doi: 10.1038/ki.1982.198

Byku, M., and Mann, D. L. (2017). "Chapter 19 - Neuromodulation of the failing heart," in *Cardioskeletal Myopathies in Children and Young Adults*, eds J. L. Jefferies, B. C. Blaxall, J. Robbins, and J. A. Towbin (Boston, MA: Academic Press), 381–397. doi: 10.1016/B978-0-12-800040-3.00019-4

Cadnapaphornchai, M. A., Tkachenko, O., Shchekochikhin, D., and Schrier, R. W. (2014). The nephrotic syndrome: pathogenesis and treatment of edema formation and secondary complications. *Pediatr. Nephrol.* 29, 1159–1167. doi: 10.1007/s00467-013-2567-8

Chalasani, N., Gorski, J. C., Horlander, J. C., Craven, R., Hoen, H., Maya, J., et al. (2001). Effects of albumin/furosemide mixtures on responses to furosemide in hypoalbuminemic patients. *J. Am. Soc. Nephrol.* 12, 1010–1016.

Chen, H. H., Neutel, J. M., Smith, D. H., Heublein, D., and Burnett, J. C. (2016). A first-in-human trial of a novel designer natriuretic peptide ZD100 in human hypertension. *J. Am. Soc. Hypertens.* 10:e23. doi: 10.1016/j.jash.2016.03.051

Crew, R. J., Radhakrishnan, J., and Appel, G. (2004). Complications of the nephrotic syndrome and their treatment. *Clin. Nephrol.* 62, 245–259. doi: 10.5414/CNP62245

Czogalla, J., Vohra, T., Penton, D., Kirschmann, M., Craigie, E., and Loffing, J. (2016). The mineralocorticoid receptor (MR) regulates ENaC but not NCC in mice with random MR deletion. *Pflügers Arch.* 468, 849–858. doi: 10.1007/s00424-016-1798-5

Damian, D. J., McNamee, R., and Carr, M. (2016). Changes in selected metabolic parameters in patients over 65 receiving hydrochlorothiazide plus amiloride, atenolol or placebo in the MRC elderly trial. *BMC Cardiovasc. Disord.* 16:188. doi: 10.1186/s12872-016-0368-2

Danielson, L. A., Kercher, L. J., and Conrad, K. P. (2000). Impact of gender and endothelin on renal vasodilation and hyperfiltration induced by relaxin in conscious rats. *Am. J. Physiol. Regul. Integr. Comp. Physiol.* 279, R1298–R1304. doi: 10.1152/ajpregu.2000.279.4.R1298

Davison, A. M., Lambie, A. T., Verth, A. H., and Cash, J. D. (1974). Salt-poor human albumin in management of nephrotic syndrome. *Br. Med. J.* 1, 481–484. doi: 10.1136/bmj.1.5906.481

de Seigneux, S., Kim, S. W., Hemmingsen, S. C., Frøkiær, J., and Nielsen, S. (2006). Increased expression but not targeting of ENaC in adrenalectomized rats with PAN-induced nephrotic syndrome. *Am. J. Physiol. Renal Physiol.* 291, F208–F217. doi: 10.1152/ajprenal.00399.2005

Deschênes, G., and Doucet, A. (2000). Collecting duct Na^+/K^+-ATPase activity is correlated with urinary sodium excretion in rat nephrotic syndromes. *J. Am. Soc. Nephrol.* 11, 604–615.

Deschênes, G., Feraille, E., and Doucet, A. (2003). Mechanisms of oedema in nephrotic syndrome: old theories and new ideas. *Nephrol. Dial. Transplant.* 18, 454–456. doi: 10.1093/ndt/18.3.454

Deschênes, G., Gonin, S., Zolty, E., Cheval, L., Rousselot, M., Martin, P.-Y., et al. (2001a). Increased synthesis and AVP unresponsiveness of Na,K-ATPase in collecting duct from nephrotic rats. *J. Am. Soc. Nephrol.* 12, 2241–2252.

Deschênes, G., Wittner, M., Stefano, A. D., Jounier, S., and Doucet, A. (2001b). Collecting duct is a site of sodium retention in PAN nephrosis: a rationale for amiloride therapy. *J. Am. Soc. Nephrol.* 12, 598–601.

Dorhout Mees, E. J. (1996). Does it make sense to administer albumin to the patient with nephrotic oedema? *Nephrol. Dial. Transplant.* 11, 1224–1226. doi: 10.1093/ndt/11.7.1224

Doucet, A., Favre, G., and Deschênes, G. (2007). Molecular mechanism of edema formation in nephrotic syndrome: therapeutic implications. *Pediatr. Nephrol.* 22, 1983–1990. doi: 10.1007/s00467-007-0521-3

Duffy, M., Jain, S., Harrell, N., Kothari, N., and Reddi, A. S. (2015). Albumin and furosemide combination for management of edema in nephrotic syndrome: a review of clinical studies. *Cells* 4, 622–630. doi: 10.3390/cells4040622

Ellis, D. (2016). Pathophysiology, evaluation, and management of edema in childhood nephrotic syndrome. *Front. Pediatr.* 3:111. doi: 10.3389/fped.2015.00111

Fallahzadeh, M. A., Dormanesh, B., Fallahzadeh, M. K., Roozbeh, J., Fallahzadeh, M. H., and Sagheb, M. M. (2017). Acetazolamide and hydrochlorothiazide followed by furosemide versus furosemide and hydrochlorothiazide followed by furosemide for the treatment of adults with nephrotic edema: a randomized trial. *Am. J. Kidney Dis.* 69, 420–427. doi: 10.1053/j.ajkd.2016.10.022

Farquhar, M. G., and Palade, G. E. (1961). Glomerular permeability: II. Ferritin transfer across the glomerular capillary wall in nephrotic rats. *J. Exp. Med.* 114, 699–716. doi: 10.1084/jem.114.5.699

Féraille, E., Vogt, B., Rousselot, M., Barlet-Bas, C., Cheval, L., Doucet, A., et al. (1993). Mechanism of enhanced Na-K-ATPase activity in cortical collecting duct from rats with nephrotic syndrome. *J. Clin. Invest.* 91, 1295–1300. doi: 10.1172/JCI116328

Firsov, D., Gautschi, I., Merillat, A.-M., Rossier, B. C., and Schild, L. (1998). The heterotetrameric architecture of the epithelial sodium channel (ENaC). *EMBO J.* 17, 344–352. doi: 10.1093/emboj/17.2.344

Fliser, D., Zurbrüggen, I., Mutschler, E., Bischoff, I., Nussberger, J., Franek, E., et al. (1999). Coadministration of albumin and furosemide in patients with the nephrotic syndrome. *Kidney Int.* 55, 629–634. doi: 10.1046/j.1523-1755.1999.00298.x

Fodor, P., Saitúa, M. T., Rodriguez, E., González, B., and Schlesinger, L. (1982). T-cell dysfunction in minimal-change nephrotic syndrome of childhood. *Am. J. Dis. Child.* 136, 713–717. doi: 10.1001/archpedi.1982.03970440057016

Gaillard, E. A., Kota, P., Gentzsch, M., Dokholyan, N. V., Stutts, M. J., and Tarran, R. (2010). Regulation of the epithelial Na^+ channel and airway surface liquid volume by serine proteases. *Pflugers Arch.* 460, 1–17. doi: 10.1007/s00424-010-0827-z

Garty, H., and Palmer, L. G. (1997). Epithelial sodium channels: function, structure, and regulation. *Physiol. Rev.* 77, 359–396. doi: 10.1152/physrev.1997.77.2.359

Garvin, J., and Sanders, K. (1991). Endothelin inhibits fluid and bicarbonate transport in part by reducing Na^+/K^+ ATPase activity in the rat proximal straight tubule. *J. Am. Soc. Nephrol.* 2, 976–982.

Geers, A. B., Koomans, H. A., Boer, P., and Dorhout Mees, E. J. (1984). Plasma and blood volumes in patients with the nephrotic syndrome. *Nephron* 38, 170–173. doi: 10.1159/000183302

Granero, G. E., Longhi, M. R., Mora, M. J., Junginger, H. E., Midha, K. K., Shah, V. P., et al. (2010). Biowaiver monographs for immediate release solid oral dosage forms: furosemide. *J. Pharm. Sci.* 99, 2544–2556. doi: 10.1002/jps.22030

Guan, Y., Hao, C., Cha, D. R., Rao, R., Lu, W., Kohan, D. E., et al. (2005). Thiazolidinediones expand body fluid volume through PPARγ stimulation of ENaC-mediated renal salt absorption. *Nat. Med.* 11, 861–866. doi: 10.1038/nm1278

Hammarlund, M. M., Paalzow, L. K., and Odlind, B. (1984). Pharmacokinetics of furosemide in man after intravenous and oral administration. Application of moment analysis. *Eur. J. Clin. Pharmacol.* 26, 197–207. doi: 10.1007/BF00630286

Haws, R. M., and Baum, M. (1993). Efficacy of albumin and diuretic therapy in children with nephrotic syndrome. *Pediatrics* 91, 1142–1146.

Hinrichs, G. R., Mortensen, L. A., Jensen, B. L., and Bistrup, C. (2018). Amiloride resolves resistant edema and hypertension in a patient with nephrotic syndrome; a case report. *Physiol. Rep.* 6:e13743. doi: 10.14814/phy2.13743

Hoorn, E. J., and Ellison, D. H. (2017). Diuretic resistance. *Am. J. Kidney Dis.* 69, 136–142. doi: 10.1053/j.ajkd.2016.08.027

Hughey, R. P., Mueller, G. M., Bruns, J. B., Kinlough, C. L., Poland, P. A., Harkleroad, K. L., et al. (2003). Maturation of the epithelial Na^+ channel involves proteolytic processing of the alpha- and gamma-subunits. *J. Biol. Chem.* 278, 37073–37082. doi: 10.1074/jbc.M307003200

Hull, R. P., and Goldsmith, D. J. (2008). Nephrotic syndrome in adults. *BMJ* 336, 1185–1189. doi: 10.1136/bmj.39576.709711.80

Humphreys, M. H. (1994). Mechanisms and management of nephrotic edema. *Kidney Int.* 45, 266–281. doi: 10.1038/ki.1994.33

Hung, S.-C., Kuo, K.-L., Peng, C.-H., Wu, C.-H., Lien, Y.-C., Wang, Y.-C., et al. (2014). Volume overload correlates with cardiovascular risk factors in patients with chronic kidney disease. *Kidney Int.* 85, 703–709. doi: 10.1038/ki.2013.336

Ichikawa, I., Rennke, H. G., Hoyer, J. R., Badr, K. F., Schor, N., Troy, J. L., et al. (1983). Role for intrarenal mechanisms in the impaired salt excretion of experimental nephrotic syndrome. *J. Clin. Invest.* 71, 91–103. doi: 10.1172/JCI110756

Inoue, M., Okajima, K., Itoh, K., Ando, Y., Watanabe, N., Yasaka, T., et al. (1987). Mechanism of furosemide resistance in analbuminemic rats and hypoalbuminemic patients. *Kidney Int.* 32, 198–203. doi: 10.1038/ki.1987.192

Iwai, N., and Inagami, T. (1990). Molecular cloning of a complementary DNA to rat cyclophilin-like protein mRNA. *Kidney Int.* 37, 1460–1465. doi: 10.1038/ki.1990.136

Johnson, M. D., and Malvin, R. L. (1977). Stimulation of renal sodium reabsorption by angiotensin II. *Am. J. Physiol. Renal Physiol.* 232, F298–F306. doi: 10.1152/ajprenal.1977.232.4.F298

Keenswijk, W., Ilias, M. I., Raes, A., Donckerwolcke, R., and Walle, J. V. (2018). Urinary potassium to urinary potassium plus sodium ratio can accurately identify hypovolemia in nephrotic syndrome: a provisional study. *Eur. J. Pediatr.* 177, 79–84. doi: 10.1007/s00431-017-3029-2

Kim, S. W., de Seigneux, S., Sassen, M. C., Lee, J., Kim, J., Knepper, M. A., et al. (2006). Increased apical targeting of renal ENaC subunits and decreased expression of 11βHSD2 in HgCl2-induced nephrotic syndrome in rats. *Am. J. Physiol. Renal Physiol.* 290, F674–F687. doi: 10.1152/ajprenal.00084.2005

Kirchner, K. A., Voelker, J. R., and Brater, D. C. (1990). Intratubular albumin blunts the response to furosemide-A mechanism for diuretic resistance in the nephrotic syndrome. *J. Pharmacol. Exp. Ther.* 252, 1097–1101.

Kleyman, T. R., Carattino, M. D., and Hughey, R. P. (2009). ENaC at the cutting edge: regulation of epithelial sodium channels by proteases. *J. Biol. Chem.* 284, 20447–20451. doi: 10.1074/jbc.R800083200

Klisic, J., Zhang, J., Nief, V., Reyes, L., Moe, O. W., and Ambühl, P. M. (2003). Albumin regulates the Na^+/H^+ exchanger 3 in OKP cells. *J. Am. Soc. Nephrol.* 14, 3008–3016. doi: 10.1097/01.ASN.0000098700.70804.D3

Koepke, J. P., and DiBona, G. F. (1987). Blunted natriuresis to atrial natriuretic peptide in chronic sodium-retaining disorders. *Am. J. Physiol.* 252, F865–F871. doi: 10.1152/ajprenal.1987.252.5.F865

Koot, B. G. P., Houwen, R., Pot, D.-J., and Nauta, J. (2004). Congenital analbuminaemia: biochemical and clinical implications. A case report and literature review. *Eur. J. Pediatr.* 163, 664–670. doi: 10.1007/s00431-004-1492-z

Korbet, S. M., Genchi, R. M., Borok, R. Z., and Schwartz, M. M. (1996). The racial prevalence of glomerular lesions in nephrotic adults. *Am. J. Kidney Dis.* 27, 647–651. doi: 10.1016/S0272-6386(96)90098-0

Kottke, M. A., and Walters, T. J. (2016). Where's the leak in vascular barriers? A review. *Shock* 46, 20–36. doi: 10.1097/SHK.0000000000000666

Lagrue, G., Branellec, A., Blanc, C., Xheneumont, S., Beaudoux, F., Sobel, A., et al. (1975a). A vascular permeability factor in lymphocyte culture supernants from patients with nephrotic syndrome. II. Pharmacological and physicochemical properties. *Biomed. Publiee Pour AAICIG* 23, 73–75.

Lagrue, G., Xheneumont, S., Branellec, A., Hirbec, G., and Weil, B. (1975b). A vascular permeability factor elaborated from lymphocytes. I. Demonstration in patients with nephrotic syndrome. *Biomedicine* 23, 37–40.

Lewis, D. M., Tooke, J. E., Beaman, M., Gamble, J., and Shore, A. C. (1998). Peripheral microvascular parameters in the nephrotic syndrome. *Kidney Int.* 54, 1261–1266. doi: 10.1046/j.1523-1755.1998.00100.x

Light, D. B., Schwiebert, E. M., Karlson, K. H., and Stanton, B. A. (1989). Atrial natriuretic peptide inhibits a cation channel in renal inner medullary collecting duct cells. *Science* 243, 383–385. doi: 10.1126/science.2463673

Loon, N. R., Wilcox, C. S., and Unwin, R. J. (1989). Mechanism of impaired natriuretic response to furosemide during prolonged therapy. *Kidney Int.* 36, 682–689. doi: 10.1038/ki.1989.246

Lourdel, S., Loffing, J., Favre, G., Paulais, M., Nissant, A., Fakitsas, P., et al. (2005). Hyperaldosteronemia and activation of the epithelial sodium channel are not required for sodium retention in puromycin-induced nephrosis. *J. Am. Soc. Nephrol.* 16, 3642–3650. doi: 10.1681/ASN.2005040363

Maack, T. (1980). Physiological evaluation of the isolated perfused rat kidney. *Am. J. Physiol.* 238, F71–F78. doi: 10.1152/ajprenal.1980.238.2.F71

Mariano, L. N. B., Boeing, T., da Silva, R. C. M. V. A. F., Cechinel-Filho, V., Niero, R., da Silva, L. M., et al. (2018). Preclinical evaluation of the diuretic and saluretic effects of (-)-epicatechin and the result of its combination with standard diuretics. *Biomed. Pharmacother.* 107, 520–525. doi: 10.1016/j.biopha.2018.08.045

Marino, F., Martorano, C., Tripepi, R., Bellantoni, M., Tripepi, G., Mallamaci, F., et al. (2016). Subclinical pulmonary congestion is prevalent in nephrotic syndrome. *Kidney Int.* 89, 421–428. doi: 10.1038/ki.2015.279

McGrogan, A., Franssen, C. F., and de Vries, C. S. (2011). The incidence of primary glomerulonephritis worldwide: a systematic review of the literature. *Nephrol. Dial. Transplant.* 26, 414–430. doi: 10.1093/ndt/gfq665

Meltzer, J. I., Keim, H. J., Laragh, J. H., Sealey, J. E., Jan, K. M., and Chien, S. (1979). Nephrotic syndrome: vasoconstriction and hypervolemic types indicated by renin-sodium profiling. *Ann. Intern. Med.* 91, 688–696. doi: 10.7326/0003-4819-91-5-688

Mernissi, G. E., and Doucet, A. (1983). Short-term effect of aldosterone on renal sodium transport and tubular Na-K-ATPase in the rat. *Pflügers Arch.* 399, 139–146. doi: 10.1007/BF00663910

Metra, M., Cotter, G., Davison, B. A., Felker, G. M., Filippatos, G., Greenberg, B. H., et al. (2013). Effect of serelaxin on cardiac, renal, and hepatic biomarkers in the Relaxin in Acute Heart Failure (RELAX-AHF) development program: correlation with outcomes. *J. Am. Coll. Cardiol.* 61, 196–206. doi: 10.1016/j.jacc.2012.11.005

Miura, T., Watanabe, S., Urushihara, M., Kobori, H., and Fukuda, M. (2014). The natriuretic effect of angiotensin receptor blockers is not attributable to blood pressure reduction during the previous night, but to inhibition of tubular sodium reabsorption. *J. Renin Angiotensin Aldosterone Syst.* 15, 316–318. doi: 10.1177/1470320313518253

Morales, E., Caro, J., Gutierrez, E., Sevillano, A., Auñón, P., Fernandez, C., et al. (2015). Diverse diuretics regimens differentially enhance the antialbuminuric effect of renin–angiotensin blockers in patients with chronic kidney disease. *Kidney Int.* 88, 1434–1441. doi: 10.1038/ki.2015.249

Nielsen, L. H., Melero, A., Keller, S. S., Jacobsen, J., Garrigues, T., Rades, T., et al. (2016). Polymeric microcontainers improve oral bioavailability of furosemide. *Int. J. Pharm.* 504, 98–109. doi: 10.1016/j.ijpharm.2016.03.050

Odlind, B. O., and Beermann, B. (1980). Diuretic resistance: reduced bioavailability and effect of oral frusemide. *Br. Med. J.* 280:1577. doi: 10.1136/bmj.280.6231.1577

Ogi, M., Yokoyama, H., Tomosugi, N., Hisada, Y., Ohta, S., Takaeda, M., et al. (1994). Risk factors for infection and immunoglobulin replacement therapy in adult nephrotic syndrome. *Am. J. Kidney Dis.* 24, 427–436. doi: 10.1016/S0272-6386(12)80899-7

Oh, S. W., and Han, S. Y. (2015). Loop diuretics in clinical practice. *Electrolytes Blood Press.* 13, 17–21. doi: 10.5049/EBP.2015.13.1.17

Orisio, S., Perico, N., Benatti, L., Longaretti, L., Amuchastegui, S., and Remuzzi, G. (1993). Renal cyclophilin-like protein gene expression parallels changes in sodium excretion in experimental nephrosis and is positively modulated by atrial natriuretic peptide. *J. Am. Soc. Nephrol.* 3, 1710–1716.

Park, E.-S., Huh, Y., and Kim, G.-H. (2015). Is tolvaptan indicated for refractory oedema in nephrotic syndrome? *Nephrology* 20, 103–106. doi: 10.1111/nep.12348

Pasini, A., Aceto, G., Ammenti, A., Ardissino, G., Azzolina, V., Bettinelli, A., et al. (2015). Best practice guidelines for idiopathic nephrotic syndrome: recommendations versus reality. *Pediatr. Nephrol.* 30, 91–101. doi: 10.1007/s00467-014-2903-7

Paskhina, T. S., Polyantseva, L. R., Krinskaya, A. V., Yegorova, T. P., Nartikova, V. P., and Levina, G. O. (1979). High concentrations of free kinins and kinin system components in abdominal transudate of a patient with nephrotic syndrome. *Clin. Chim. Acta* 97, 73–82. doi: 10.1016/0009-8981(79)90026-3

Passero, C. J., Hughey, R. P., and Kleyman, T. R. (2010). New role for plasmin in sodium homeostasis. *Curr. Opin. Nephrol. Hypertens.* 19, 13–19. doi: 10.1097/MNH.0b013e3283330fb2

Patel-Chamberlin, M., Kia, M. V., Xu, J., Barone, S., Zahedi, K., and Soleimani, M. (2016). The role of epithelial sodium channel ENaC and the apical Cl-/HCO3-exchanger pendrin in compensatory salt reabsorption in the setting of Na-Cl cotransporter (NCC) inactivation. *PLoS One* 11:e0150918. doi: 10.1371/journal.pone.0150918

Pavlov, T. S., Imig, J. D., and Staruschenko, A. (2010). Regulation of ENaC-mediated sodium reabsorption by peroxisome proliferator-activated receptors. *PPAR Res.* 2010:703735. doi: 10.1155/2010/703735

Perico, N., Delaini, F., Lupini, C., Benigni, A., Galbusera, M., Boccardo, P., et al. (1989). Blunted excretory response to atrial natriuretic peptide in experimental nephrosis. *Kidney Int.* 36, 57–64. doi: 10.1038/ki.1989.161

Perico, N., and Remuzzi, G. (1993). Edema of the nephrotic syndrome: the role of the atrial peptide system. *Am. J. Kidney Dis.* 22, 355–366. doi: 10.1016/S0272-6386(12)70137-3

Perrier, R., Boscardin, E., Malsure, S., Sergi, C., Maillard, M. P., Loffing, J., et al. (2016). Severe salt–losing syndrome and hyperkalemia induced by adult nephron–specific knockout of the epithelial sodium channel α-subunit. *J. Am. Soc. Nephrol.* 27, 2309–2318. doi: 10.1681/ASN.2015020154

Polzin, D., Kaminski, H. J., Kastner, C., Wang, W., Krämer, S., Gambaryan, S., et al. (2010). Decreased renal corin expression contributes to sodium retention in proteinuric kidney diseases. *Kidney Int.* 78, 650–659. doi: 10.1038/ki.2010.197

Ray, E. C. (2018). ENaC blockade in proteinuria-associated extracellular fluid volume overload – effective but risky. *Physiol. Rep.* 6:e13835. doi: 10.14814/phy2.13835

Ray, E. C., Miller, R. G., Demko, J. E., Costacou, T., Kinlough, C. L., Demko, C. L., et al. (2018). Urinary plasmin(ogen) as a prognostic factor for hypertension. *Kidney Int. Rep.* 3, 1434–1442. doi: 10.1016/j.ekir.2018.06.007

Ray, E. C., Rondon-Berrios, H., Boyd, C. R., and Kleyman, T. R. (2015). Sodium retention and volume expansion in nephrotic syndrome: implications for hypertension. *Adv. Chronic Kidney Dis.* 22, 179–184. doi: 10.1053/j.ackd.2014.11.006

Reiser, J. (2013). Circulating permeability factor suPAR: from concept to discovery to clinic. *Trans. Am. Clin. Climatol. Assoc.* 124, 133–138.

Rossier, B. C., and Stutts, M. J. (2009). Activation of the epithelial sodium channel (ENaC) by serine proteases. *Annu. Rev. Physiol.* 71, 361–379. doi: 10.1146/annurev.physiol.010908.163108

Rostoker, G., Behar, A., and Lagrue, G. (2000). Vascular hyperpermeability in nephrotic edema. *Nephron* 85, 194–200. doi: 10.1159/000045661

Roustit, M., Genty, C., Lepelley, M., Blaise, S., Fromy, B., Cracowski, J.-L., et al. (2016). Amiloride treatment and increased risk of pressure ulcers in hospitalized patients. *Br. J. Clin. Pharmacol.* 82, 1685–1687. doi: 10.1111/bcp.13084

Sahu, B. P., and Das, M. K. (2014). Formulation, optimization, and in vitro/in vivo evaluation of furosemide nanosuspension for enhancement of its oral bioavailability. *J. Nanoparticle Res.* 16:2360. doi: 10.1007/s11051-014-2360-z

Savin, V. J., Sharma, M., Zhou, J., Genochi, D., Sharma, R., Srivastava, T., et al. (2017). Multiple targets for novel therapy of FSGS associated with circulating permeability factor. *Biomed Res. Int.* 2017:6232616. doi: 10.1155/2017/6232616

Schneider, M. P., Boesen, E. I., and Pollock, D. M. (2007). Contrasting actions of endothelin ETA and ETB receptors in cardiovascular disease. *Annu. Rev. Pharmacol. Toxicol.* 47, 731–759. doi: 10.1146/annurev.pharmtox.47.120505.105134

Sheng, S., Carattino, M. D., Bruns, J. B., Hughey, R. P., and Kleyman, T. R. (2006). Furin cleavage activates the epithelial Na^+ channel by relieving Na^+ self-inhibition. *Am. J. Physiol. Renal Physiol.* 290, F1488–F1496. doi: 10.1152/ajprenal.00439.2005

Shimizu, M., Ishikawa, S., Yachi, Y., Muraoka, M., Tasaki, Y., Iwasaki, H., et al. (2014). Tolvaptan therapy for massive edema in a patient with nephrotic syndrome. *Pediatr. Nephrol.* 29, 915–917. doi: 10.1007/s00467-013-2687-1

Sica, D. A. (2003). Drug absorption in the management of congestive heart failure: loop diuretics. *Congest. Heart Fail.* 9, 287–292. doi: 10.1111/j.1527-5299.2003.02399.x

Siddall, E. C., and Radhakrishnan, J. (2012). The pathophysiology of edema formation in the nephrotic syndrome. *Kidney Int.* 82, 635–642. doi: 10.1038/ki.2012.180

Soleimani, M., Barone, S., Xu, J., Shull, G. E., Siddiqui, F., Zahedi, K., et al. (2012). Double knockout of pendrin and Na-Cl cotransporter (NCC) causes severe salt wasting, volume depletion, and renal failure. *Proc. Natl. Acad. Sci. U.S.A.* 109, 13368–13373. doi: 10.1073/pnas.1202671109

Stæhr, M., Buhl, K. B., Andersen, R. F., Svenningsen, P., Nielsen, F., Hinrichs, G. R., et al. (2015). Aberrant glomerular filtration of urokinase-type plasminogen activator in nephrotic syndrome leads to amiloride-sensitive plasminogen activation in urine. *Am. J. Physiol. Renal Physiol.* 309, F235–F241. doi: 10.1152/ajprenal.00138.2015

Svenningsen, P., Bistrup, C., Friis, U. G., Bertog, M., Haerteis, S., Krueger, B., et al. (2009). Plasmin in nephrotic urine activates the epithelial sodium channel. *J. Am. Soc. Nephrol.* 20, 299–310. doi: 10.1681/ASN.2008040364

Takamura, N., Maruyama, T., Ahmed, S., Suenaga, A., and Otagiri, M. (1997). Interactions of aldosterone antagonist diuretics with human serum proteins. *Pharm. Res.* 14, 522–526. doi: 10.1023/A:1012168020545

Tanaka, A., Nakamura, T., Sato, E., Ueda, Y., and Node, K. (2017). Different effects of tolvaptan in patients with idiopathic membranous nephropathy with nephrotic syndrome. *Intern. Med.* 56, 191–196. doi: 10.2169/internalmedicine.56.7539

Teerlink, J. R. (2017). *RELAXin in Acute Heart Failure-2*. Paris: ACC.

Teerlink, J. R., Voors, A. A., Ponikowski, P., Pang, P. S., Greenberg, B. H., Filippatos, G., et al. (2017). Serelaxin in addition to standard therapy in acute heart failure: rationale and design of the RELAX-AHF-2 study. *Eur. J. Heart Fail.* 19, 800–809. doi: 10.1002/ejhf.830

Terker, A. S., Zhang, C., Erspamer, K. J., Gamba, G., Yang, C.-L., and Ellison, D. H. (2016). Unique chloride-sensing properties of WNK4 permit the distal nephron to modulate potassium homeostasis. *Kidney Int.* 89, 127–134. doi: 10.1038/ki.2015.289

The National Institute for Health, and Care [NICE] (2018). *BNF: British National Formulary*. London: NICE.

Trimarchi, H., Forrester, M., Lombi, F., Pomeranz, V., Raña, M. S., Karl, A., et al. (2014). Amiloride as an alternate adjuvant antiproteinuric agent in fabry disease: the potential roles of plasmin and uPAR. *Case Rep. Nephrol.* 2014:854521. doi: 10.1155/2014/854521

Udwan, K., Brideau, G., Fila, M., Edwards, A., Vogt, B., and Doucet, A. (2016). Oxidative stress and nuclear factor κB (NF-κB) increase peritoneal filtration and contribute to ascites formation in nephrotic syndrome. *J. Biol. Chem.* 291, 11105–11113. doi: 10.1074/jbc.M116.724690

Ueda, K., Hirahashi, J., Seki, G., Tanaka, M., Kushida, N., Takeshima, Y., et al. (2014). Successful treatment of acute kidney injury in patients with idiopathic nephrotic syndrome using human atrial natriuretic Peptide. *Intern. Med.* 53, 865–869. doi: 10.2169/internalmedicine.53.1724

Unruh, M. L., Pankratz, V. S., Demko, J. E., Ray, E. C., Hughey, R. P., and Kleyman, T. R. (2017). Trial of amiloride in type 2 diabetes with proteinuria. *Kidney Int. Rep.* 2, 893–904. doi: 10.1016/j.ekir.2017.05.008

Valentin, J. P., Qiu, C., Muldowney, W. P., Ying, W. Z., Gardner, D. G., and Humphreys, M. H. (1992). Cellular basis for blunted volume expansion natriuresis in experimental nephrotic syndrome. *J. Clin. Invest.* 90, 1302–1312. doi: 10.1172/JCI115995

Vassalli, J.-D., and Belin, D. (1987). Amiloride selectively inhibits the urokinase-type plasminogen activator. *FEBS Lett.* 214, 187–191. doi: 10.1016/0014-5793(87)80039-X

Voors, A. A., Davison, B. A., Teerlink, J. R., Felker, G. M., Cotter, G., Filippatos, G., et al. (2014). Diuretic response in patients with acute decompensated heart failure: characteristics and clinical outcome—an analysis from RELAX-AHF. *Eur. J. Heart Fail.* 16, 1230–1240. doi: 10.1002/ejhf.170

Waller, E. S., Hamilton, S. F., Massarella, J. W., Sharanevych, M. A., Smith, R. V., Yakatan, G. J., et al. (1981). Disposition and absolute bioavailability of furosemide in healthy males. *J. Pharm. Sci.* 71, 1105–1108. doi: 10.1002/jps.2600711006

Wang, T., and Giebisch, G. (1996). Effects of angiotensin II on electrolyte transport in the early and late distal tubule in rat kidney. *Am. J. Physiol. Renal Physiol.* 271, F143–F149. doi: 10.1152/ajprenal.1996.271.1.F143

Wang, W., Li, C., Nejsum, L. N., Li, H., Kim, S. W., Kwon, T.-H., et al. (2006). Biphasic effects of ANP infusion in conscious, euvolemic rats: roles of AQP2 and ENaC trafficking. *Am. J. Physiol. Renal Physiol.* 290, F530–F541.

Wang, Y., Bu, J., Zhang, Q., Chen, K., Zhang, J., and Bao, X. (2015). Expression pattern of aquaporins in patients with primary nephrotic syndrome with edema. *Mol. Med. Rep.* 12, 5625–5632. doi: 10.3892/mmr.2015.4209

Warnock, D. G. (2015). Amiloride: the "new" renal tonic? *Am. J. Physiol. Renal Physiol.* 309, F429–F430. doi: 10.1152/ajprenal.00237.2015

Webb, N. J. A., Watson, C. J., Roberts, I. S., Bottomley, M. J., Jones, C. A., Lewis, M. A., et al. (1999). Circulating vascular endothelial growth factor is not increased during relapses of steroid-sensitive nephrotic syndrome. *Kidney Int.* 55, 1063–1071. doi: 10.1046/j.1523-1755.1999.055003 1063.x

Wilcox, C. S., Mitch, W. E., Kelly, R. A., Skorecki, K., Meyer, T. W., Friedman, P. A., et al. (1983). Response of the kidney to furosemide: I. Effects of salt intake and renal compensation. *J. Lab. Clin. Med.* 102, 450–458.

Wile, D. (2012). Diuretics: a review. *Ann. Clin. Biochem.* 49, 419–431. doi: 10.1258/acb.2011.011281

Wilson, S. S., Ayaz, S. I., and Levy, P. D. (2015). Relaxin: a novel agent for the treatment of acute heart failure. *Pharmacother. J. Hum. Pharmacol. Drug Ther.* 35, 315–327. doi: 10.1002/phar.1548

Zahedi, K., Barone, S., Xu, J., and Soleimani, M. (2013). Potentiation of the effect of thiazide derivatives by carbonic anhydrase inhibitors: molecular mechanisms and potential clinical implications. *PLoS One* 8:e79327. doi: 10.1371/journal.pone.0079327

Zhang, B., Xie, S., Shi, W., and Yang, Y. (2012). Amiloride off-target effect inhibits podocyte urokinase receptor expression and reduces proteinuria. *Nephrol. Dial. Transplant.* 27, 1746–1755. doi: 10.1093/ndt/gfr612

Zhang, G., Kernan, K. A., Collins, S. J., Cai, X., López-Guisa, J. M., Degen, J. L., et al. (2007). Plasmin(ogen) promotes renal interstitial fibrosis by promoting epithelial-to-mesenchymal transition: role of plasmin-activated signals. *J. Am. Soc. Nephrol.* 18, 846–859. doi: 10.1681/ASN.2006080886

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Nephrol Dial Transplant (2017) 1–9
doi: 10.1093/ndt/gfx296

ndt
Nephrology Dialysis Transplantation

REVIEW

Genetics of membranous nephropathy

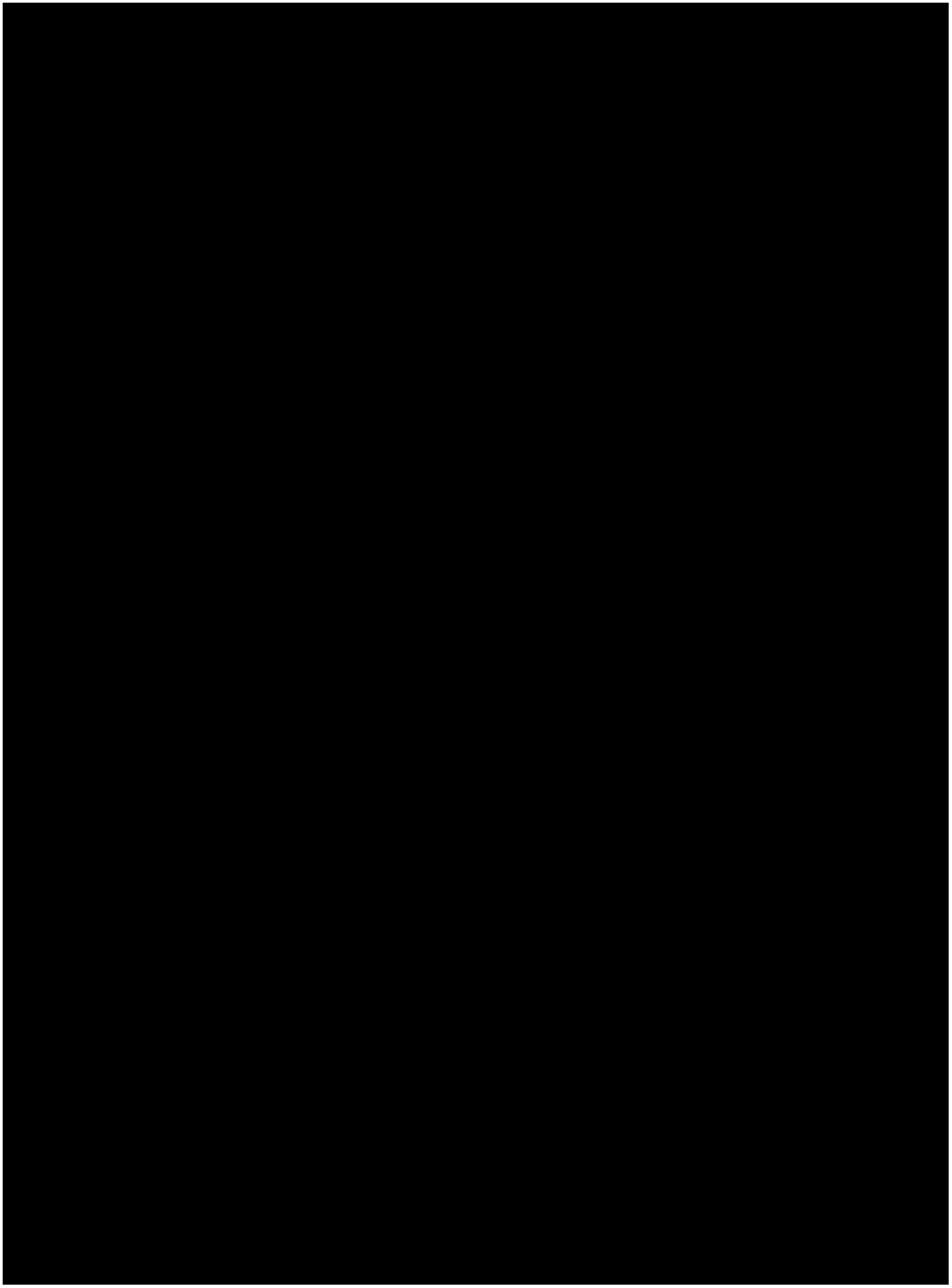
Sanjana Gupta[1], Anna Köttgen[2], Elion Hoxha[3], Paul Brenchley[4], Detlef Bockenhauer[1], Horia C. Stanescu[1] and Robert Kleta[1]

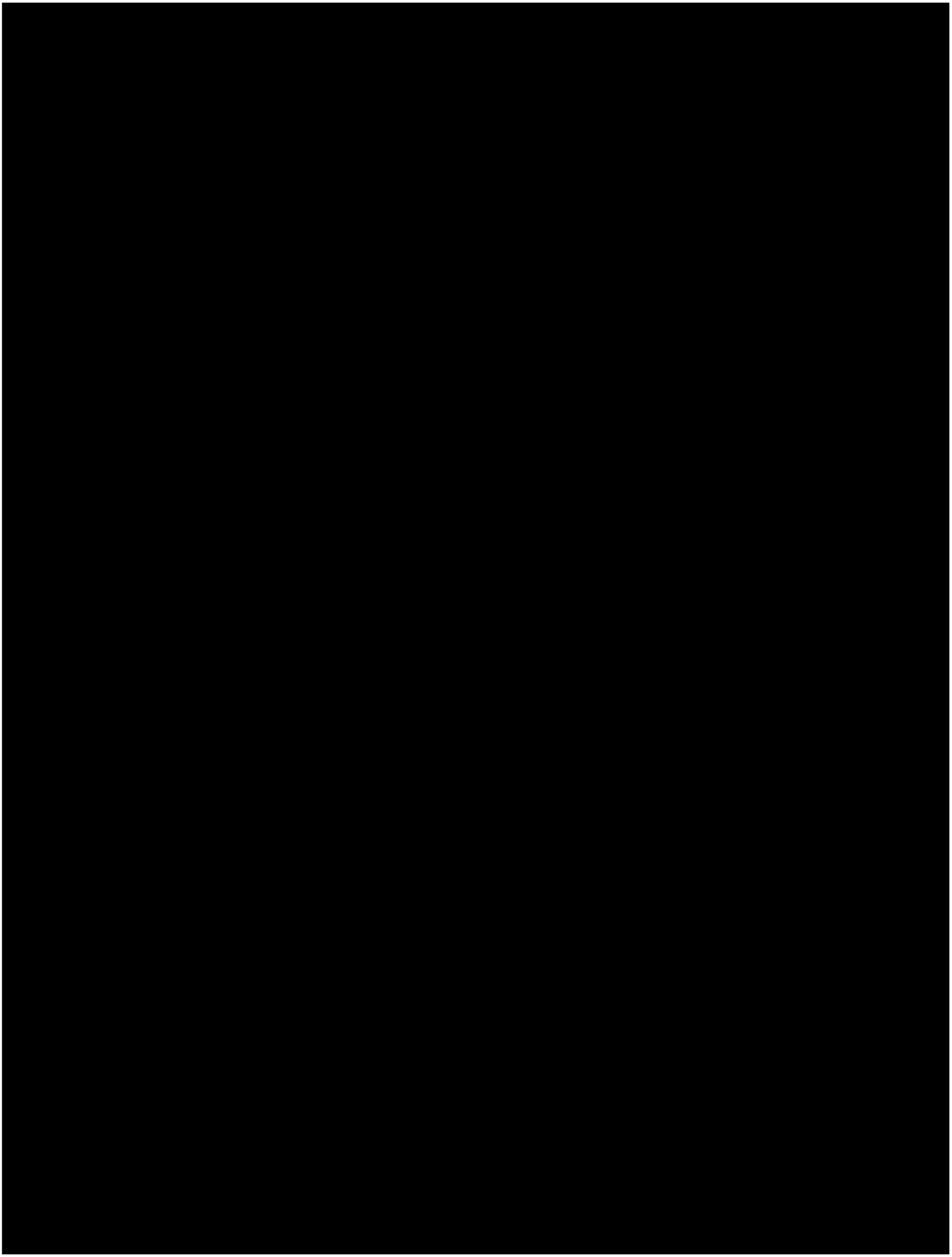
[1]University College London–Centre for Nephrology, London, UK, [2]Institute of Genetic Epidemiology, Medical Center and Faculty of Medicine, University of Freiburg, Freiburg, Germany, [3]Medizinische Klinik und Poliklinik III, Universitätsklinikum Hamburg-Eppendorf, Hamburg, Germany and [4]Institute of Cardiovascular Sciences, University of Manchester, Manchester, UK

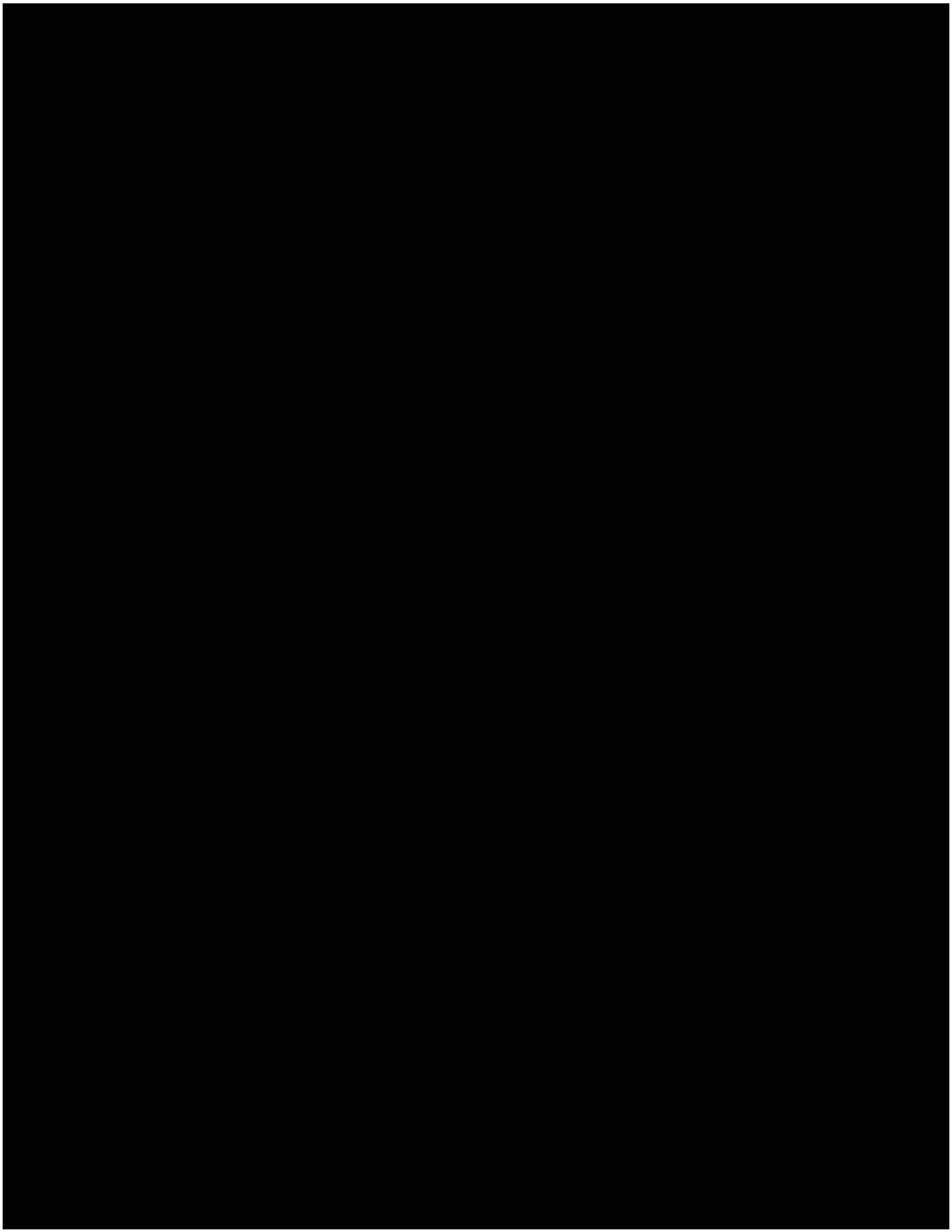
Correspondence and offprint requests to: Robert Kleta; E-mail r.kleta@ucl.ac.uk

ABSTRACT

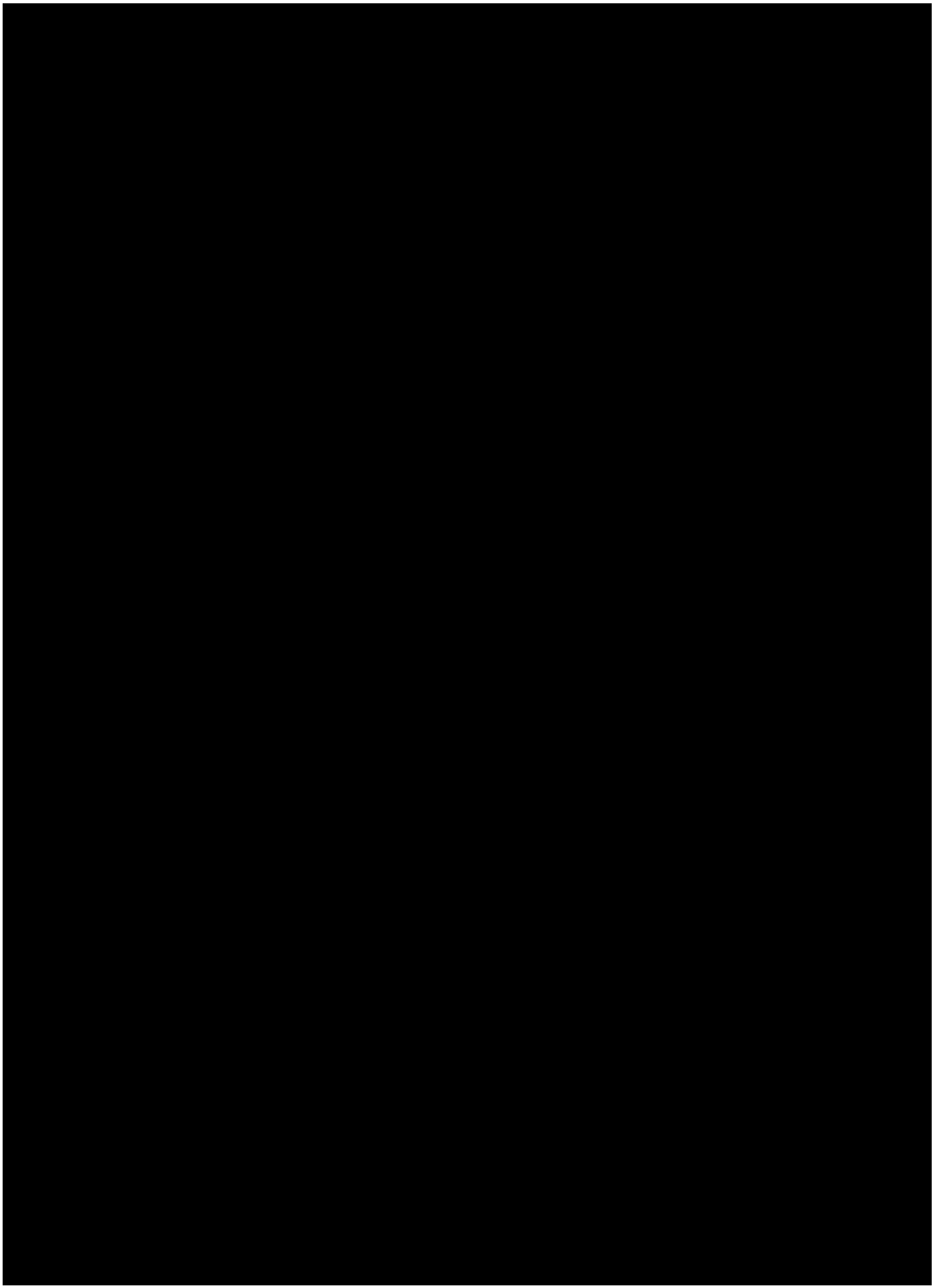
An HLA-DR3 association with membranous nephropathy (MN) was described in 1979 and additional evidence for a genetic component to MN was suggested in 1984 in reports of familial MN. In 2009, a pathogenic autoantibody was identified against the phospholipase A_2 receptor 1 (PLA_2R1). Here we discuss the genetic studies that have proven the association of human leucocyte antigen class II and PLA_2R1 variants and disease in MN. The common variants in PLA_2R1 form a haplotype that is associated with disease incidence. The combination of the variants in both genes significantly increases the risk of disease by 78.5-fold. There are important genetic ethnic differences in MN. Disease outcome is difficult to predict and attempts to correlate the genetic association to outcome have so far not been helpful in a reproducible manner. The role of genetic variants may not only extend beyond the risk of disease development, but can also help us understand the underlying molecular biology of the PLA_2R1 and its resultant pathogenicity. The genetic variants identified thus far have an association with disease and could therefore become useful biomarkers to stratify disease risk, as well as possibly identifying novel drug targets in the near future.

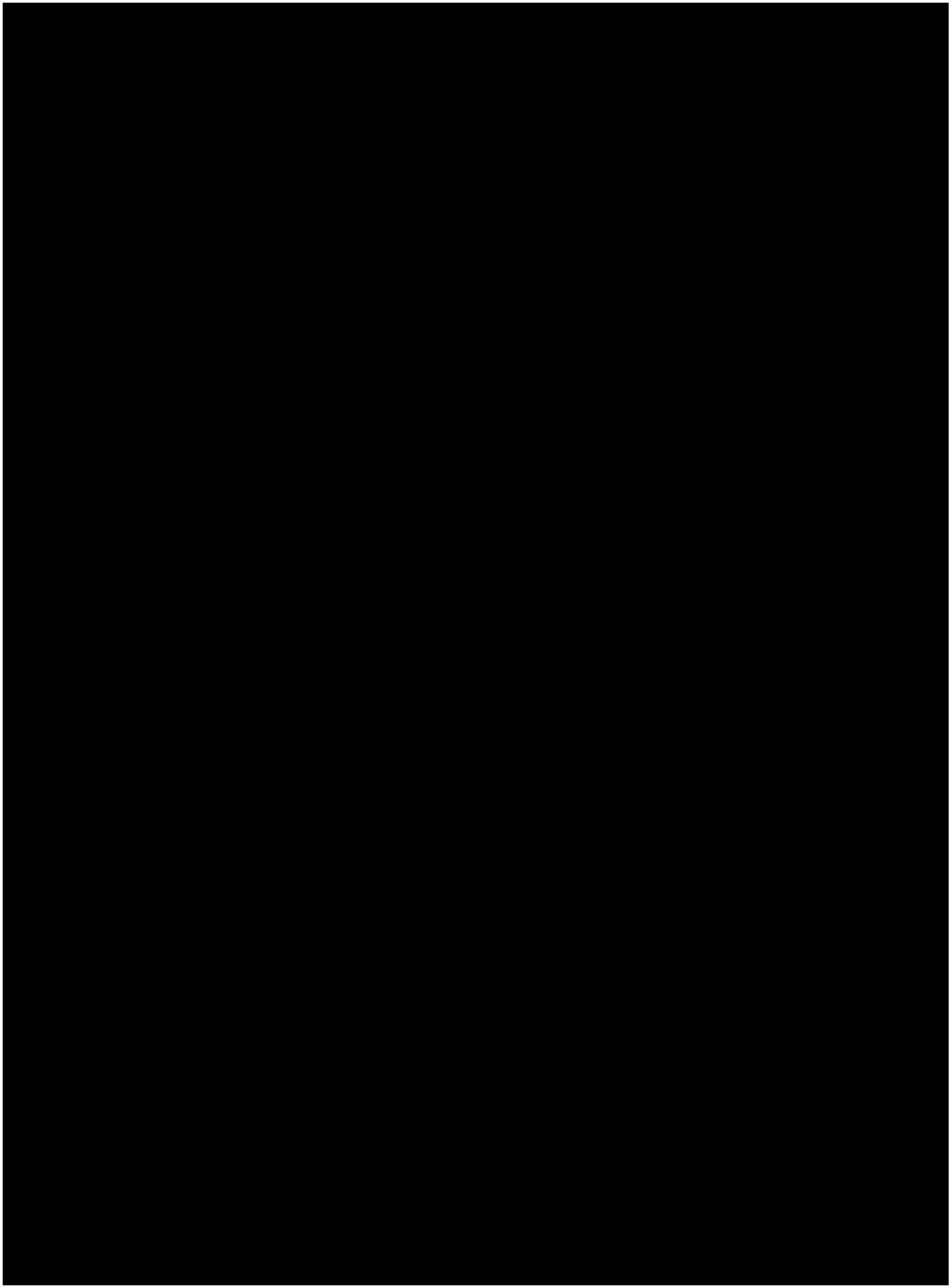


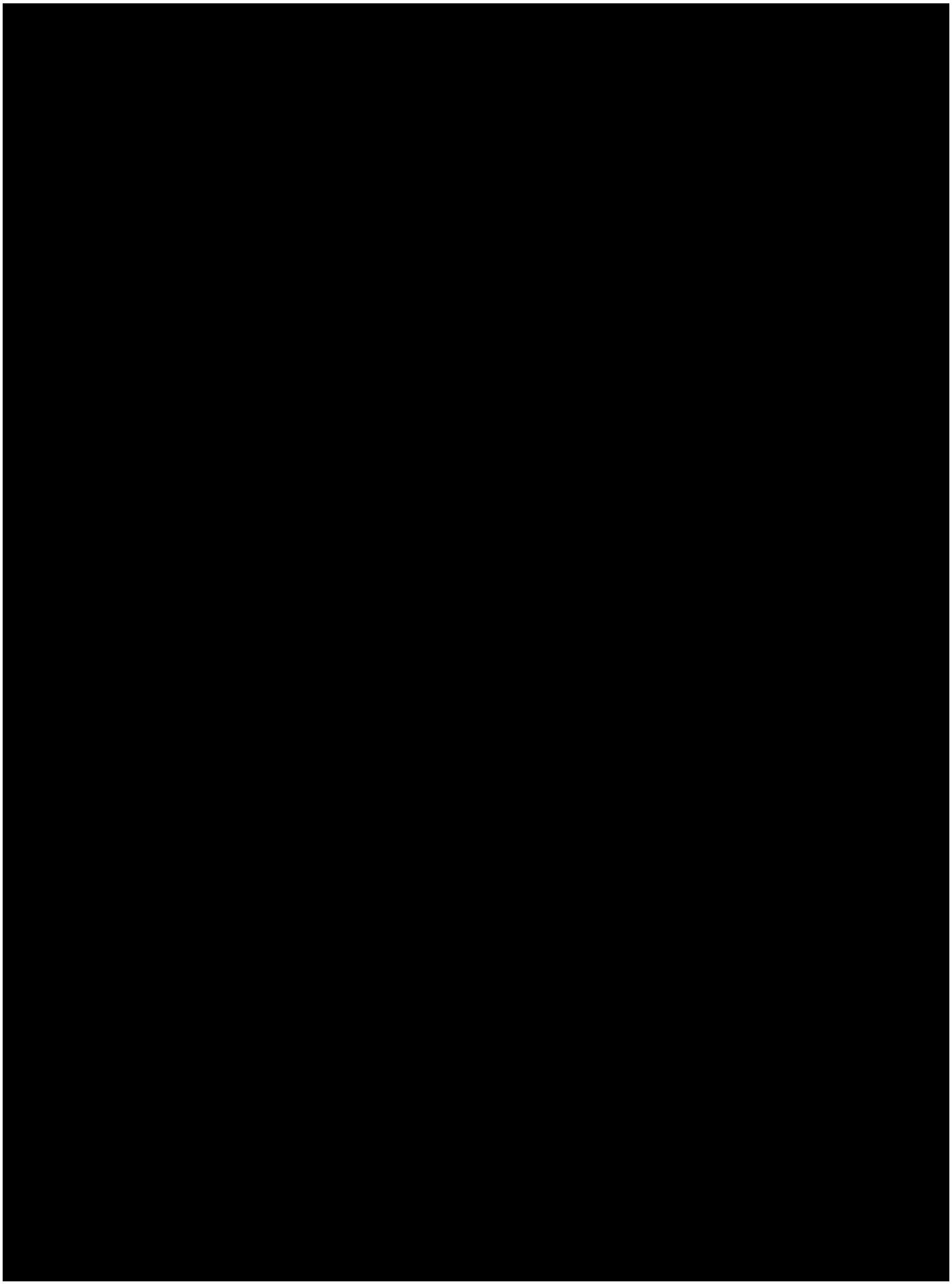




Continued







Gupta et al. BMC Nephrology (2017) 18:201
DOI 10.1186/s12882-017-0615-5

BMC Nephrology

RESEARCH ARTICLE Open Access

Membranous nephropathy: a retrospective observational study of membranous nephropathy in north east and central London

Sanjana Gupta[1,2], John Connolly[1], Ruth J Pepper[1], Stephen B Walsh[1], Magdi M Yaqoob[2], Robert Kleta[1*] and Neil Ashman[2]

Abstract

Background: Membranous nephropathy (MN) is the leading cause of nephrotic syndrome in adults. MN is a clinically heterogeneous disease and it is difficult to accurately predict outcomes (including end stage renal failure) at presentation and whom to treat with potentially toxic therapies. We aimed to identify factors predicting outcome in MN in our cohort from two large tertiary London units by undertaking a retrospective data analysis of 148 biopsy-proven MN patients from North East and Central London between 1995 and 2015.

Methods: Review of clinical and biochemistry databases.

Results: Surprisingly, patients that reached end stage renal failure (ESRF) had a less severe nephrosis compared to those that did not develop ESRF; serum albumin 33 g/L (3.3 g/dL) versus 24 g/L (2.4 g/dL), $p = 0.002$ and urinary protein creatinine ratio (uPCR) 550 mg/mmol (5500 mg/g) versus 902 mg/mmol (9020 mg/g), $p = 0.0124$. The correlation with ESRF was strongest with the presenting creatinine; 215 μmol/L (2.43 mg/dL) compared to 81 μmol/L (0.92 mg/dL), $p < 0.0001$. Patients presenting with creatinine of >120 μmol/L (1.36 mg/dL; corresponding to an eGFR of ≤60 ml/min in non-Black males) had an increased rate of ESRF and a faster decline. Other traditional risk factors for progression were not significantly associated with ESRF.
Black patients presented with higher serum creatinine but no statistically significant difference in the estimated glomerular filtration rate, a higher rate of progression to ESRF and had a poorer response to treatment.

Conclusions: This ethnically diverse cohort does not demonstrate the traditional risk profile associated with development of ESRF. Thus, careful consideration of therapeutic options is crucial, as current risk modelling cannot accurately predict the risk of ESRF. Further studies are required to elucidate the role of antibodies and risk genes.

Keywords: Membranous nephropathy, Renal failure, Ethnic differences, Nephrotic syndrome, Risk factors

Background

Idiopathic membranous nephropathy (IMN) is a serious autoimmune renal disease that is the leading cause of adult nephrotic syndrome and can progress to end stage renal failure (ESRF). Secondary forms exist that are attributable to an underlying cause. In all patients with membranous nephropathy (MN) the pathogenesis involves the development of autoantibodies against antigens present on podocytes. Classic autoimmune disorders have a strong female preponderance [1, 2], whereas with MN males are predominantly affected (with a ratio of approximately 3:1). MN has a variable natural history and tends to develop in a stratified way. It demonstrates an approximate 'rule of thirds': in untreated patients, spontaneous complete remission of proteinuria occurs in 5-30% at 5 years [3–5], spontaneous partial remission in 25-40% at 5 years [3–5] and progression to ESRF in 41% at 5 years [4, 6]. The risk of progressing to ESRF is increased in those who are older

* Correspondence: r.kleta@ucl.ac.uk
[1]UCL Centre for Nephrology, 1st Floor, Room 1.7007, Rowland Hill Street, London NW3 2PF, UK
Full list of author information is available at the end of the article

at presentation, have nephrotic range proteinuria and/or decreased glomerular filtration rate (GFR) at presentation; interestingly it is also increased in males [3, 7, 8]. Asian patients appear to have a better prognosis than non-Asians [7].

Immunomodulatory treatment for MN includes cyclophosphamide (CYC), chlorambucil, calcineurin inhibitors (CNI) - such as cyclosporine A and tacrolimus, rituximab, anti-proliferative agents (AP) - such as mycophenolate mofetil and azathioprine - and corticosteroids. These all predispose to opportunistic infections. Alkylating agents, the gold standard treatment recommended by KDIGO [9], increase cancer risk threefold [10]. To lessen exposure to these therapeutic toxins, there has been much interest in predicting MN patients at risk of progression to ESRF. The predictive accuracy of heavy proteinuria is only 30–50%, and risk modeling with multiple clinical variables (still based on data from 1997) yields a disappointing 80% accuracy rate [11].

Published studies describe ethnically homogenous patient cohorts [4, 12, 13] and therefore we were interested to see if there were differences at diagnosis, treatment or response rate within two tertiary renal London units that cover an extensive and ethnically diverse area of North East and Central London. This retrospective study was undertaken to ascertain if there are differences in our patient population, treatment strategies and remission rates compared to those previously reported.

Methods

Patient selection

Our study was performed across two tertiary London Renal Units – The Royal Free Hospital and the Royal London Hospital. We identified adult patients with membranous nephropathy (MN) by searching the clinical renal databases at both centres. We excluded patients that did not have MN. Two hundred forty patients were identified with biopsy proven MN. A further 92 patients were excluded from analysis as there was no serial data available for the 2 year period after the diagnosis of MN was made. The remaining 148 patients were included in the study. Of these 148, 121 had IMN, 4 de novo MN in renal transplants and 23 secondary MN. The patients with secondary MN had a range of causes: 14 systemic lupus erythematosus, 1 scleroderma, 6 hepatitides, 1 malignancy and 1 tuberculosis. The study was retrospective so did not need ethical approval as per NHS Health Research Authority regulation.

Data collection

Data were collected using the renal databases in addition to local clinical pathology databases. The date of the biopsy was considered to be month 0 – (date of diagnosis) and subsequent data collection based thereafter on this date. Serial data was collected at diagnosis, 1, 2, 3, 6, 12, 18 and 24 months. At each time point serum creatinine, albumin, urine protein creatinine ratio, haemoglobin, immunoglobulins and bicarbonate were recorded. Additionally, the use of renin angiotensin system blocker (RASB) or immunosuppression was recorded at each month. Rates of complications, co-morbidities as well as demographic data such as gender, age and ethnicity was collected. Remission status was calculated based on the standard criteria for complete and partial remission [13].

Analysis of results

A retrospective analysis was then performed and data analysed using Graphpad Prism 6 (Graphpad software, USA). For parametric data, t-tests were used to compare two data sets, and Mann-Whitney tests for non-parametric data. Contingency tables were analysed with Chi square tests and more than three data sets were compared with ANOVA analysis. Prism 6 was used to formulate the graphs.

Results

Baseline characteristics

A total of 148 patients were included in this retrospective study. The baseline characteristics of the study population are described in Table 1.

Differences at diagnosis

In our study population there were significant differences between characteristics at diagnosis in those that reached

Table 1 Table demonstrating baseline characteristics of all patients. Values are given either as median with interquartile range (IQR) or mean with standard deviation (SD) or percentages

Number of cases	148
Gender Ratio (Male: Female) n	90: 58
Median Age	58 (47–71)
Ethnicity (Asian: Black: White: Unknown) %	31: 24: 36: 9
Median Diagnosis Serum creatinine μmol/L (IQR) (mg/dL)	92 (68–183) (1.04, 0.77–2.07)
Median Diagnosis Serum albumin g/L (IQR) (g/dL)	25 (20–31) (2.5, 2–3.1)
Median Diagnosis urine protein creatinine ratio mg/mmol (IQR) (mg/g)	776 (432 – 1172) (7760, 4320–11,720)
Median Diagnosis cholesterol mmol/L (IQR) (mg/dL)	7.5 (5.75–9.25) (290, 222–357)
Median Diagnosis bicarbonate mmol/L (IQR)	25 (23–28)
Mean Diagnosis haemoglobin g/L (SD)	124.8 (21.68)
Co-morbidities: Hypertension / Diabetes / Recurrent UTIs / Malignancy / Mental health issues n	38 / 24 / 4 / 1 / 2
Complications: thrombotic event / treatment related side effect %	13 / 5
Renin angiotensin system blockade medication use %	84

ESRF and those that did not. The serum creatinine was significantly higher in those reaching ESRF, 215 μmol/L (124–360) (2.43 mg/dL, 1.4–4) compared to those that did not reach ESRF, 81 μmol/L (64–120) (0.92 mg/dL, 0.72–1.36), $p < 0.0001$; Fig. 1. At each time point reviewed, the difference in the serum creatinine remained statistically significant (month 1, 2, 3, 6, 12, 18 and 24); all p-values <0.0001. Serum bicarbonate was lower compared to non-ESRF patients; 21.8 mmol/L ± 0.78 versus 26.0 mmol/L ± 0.45 in non-ESRF patients, $p < 0.0001$. Finally, haemoglobin was also lower in those that reached ESRF; 111.5 g/L ± 2.5 compared to 127.7 g/L ± 2.3, $p = 0.0001$.

Patients reaching ESRF had less severe nephrotic syndrome at presentation with a higher serum albumin; median 33 g/L (27–39) (3.3 g/dL) compared to non-ESRF patients; 24 g/L (19–30) (2.4 g/dL), $p = 0.0002$ (Fig. 2). The uPCR was also lower in ESRF patients; 550 mg/mmol (213–985) (5500 mg/g, 2130–9850) compared to 902 mg/mmol (532–1314) (9020 mg/g, 5320–13,140), $p = 0.0124$. The serum cholesterol was lower in patients with ESRF and less severe nephrotic syndrome; 5.7 mmol/L (4.4–7.9) (220 mg/dL, 170–305) compared to non-ESRF patients 7.9 mmol/L (6–9.8) (305 mg/dL, 232–378), $p = 0.008$.

There was no significant difference in gender distribution between the ESRF and non-ESRF groups, the proportion of men reaching ESRF was 27% compared to 22% women, in contrast to 73% of men being non-ESRF and 78% women, $p = 0.56$. There was also no difference in the mean age at presentation (58 ± 1.7 in the non-ESRF group compared to the ESRF group 59 ± 2.4, $p = 0.67$).

Multivariate analysis with a 2-way ANOVA and Bonferroni correction demonstrated that only two significant variables were associated with developing ESRF; the diagnosis serum creatinine and uPCR. Serum creatinine was higher in those reaching ESRF with lower uPCR, p-values 0.0134 and <0.0001 respectively.

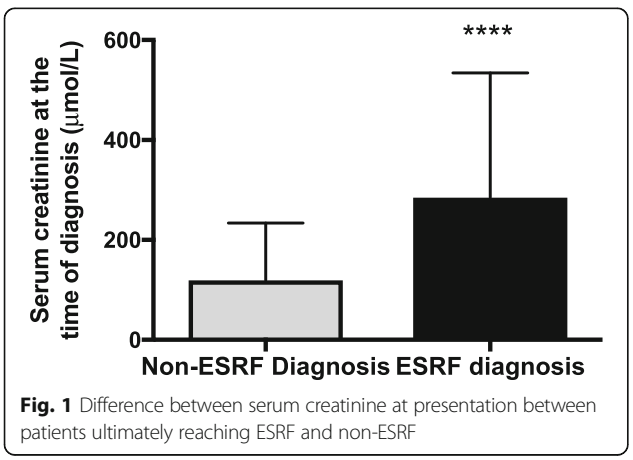


Fig. 1 Difference between serum creatinine at presentation between patients ultimately reaching ESRF and non-ESRF

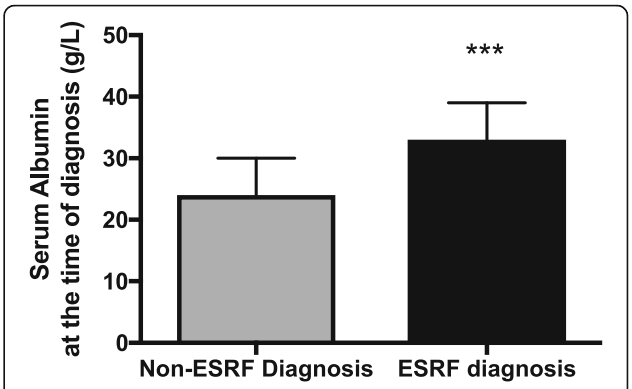


Fig. 2 Difference between serum albumin at presentation between patients reaching ESRF and non-ESRF

Progression of biochemical parameters

Detailed biochemistry was analysed for the 2 years following biopsy diagnosis in all patients. Over this period there was no significant change in the serum creatinine in the non-ESRF patients. There was also no change in the serum bicarbonate or haemoglobin. There was a significant reduction in the cholesterol over the follow up period with treatment; 7.8 mmol/L (301 mg/dL) at admission compared to 4.8 mmol/L (185 mg/dL) at 2 years, $p < 0.0001$. The albumin significantly incremented up to 41 g/L (4.1 g/dL) compared to 24 g/L (2.4 g/dL) at diagnosis, $p < 0.0001$. This mirrored a reduction in uPCR; 903 mg/mmol (9030 mg/g) at diagnosis compared to 119 mg/mmol (1190 mg/g), $p < 0.0001$.

Our patient cohort has a median follow up of 84 months, (longest 211 months / 17.5 years). We therefore examined long-term data on those reaching ESRF. A serum cut off of <120 μmol/L (1.36 mg/dL) was used as this represents an estimated GFR (eGFR) of 60 ml/min/1.73m^2 (using the abbreviated MDRD formula) in a 50 year old non-Black male. Of patients presenting with a creatinine of <120 μmol/L (1.36 mg/dL) only 10% (9 of 89) reached ESRF within a median time to ESRF of 178 months (under 15 years). Patients with a creatinine >120 μmol/L (1.36 mg/dL) however had an increased rate of developing ESRF; 29 out of 54 (54%) and at a quicker rate with a median time period of 117 months (under 10 years), this difference is statistically significant, $p < 0.0001$, Fig. 3.

Ethnicity differences

Our study population are ethnically diverse, enabling direct comparisons between different ethnic groups. There is an approximate equal spread over the different ethnic groups; Table 2. There was no significant difference in the age at diagnosis for these different groups (mean age – Asian 57, Black 57, White 60). Median creatinine was significantly higher in Black patients (103 μmol/L / 1.17 mg/dL) compared to Asian (69.5 μmol/L / 0.79 mg/dL) and White

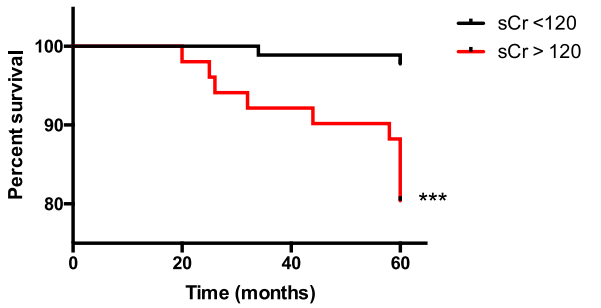


Fig. 3 Survival graph showing time to renal failure and difference between presentation creatinine and survival at 5 years

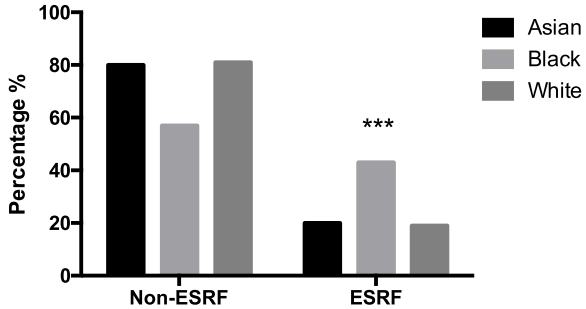


Fig. 4 Difference in ethnicities and resultant ESRF rates

(87.5 μmol/L / 0.99 mg/dL), ANOVA $p = 0.0443$. However, there was no statistically significant difference between MDRD eGFR though there was a trend to lower eGFR in Black patients. Black patients eGFR was 67 ml/min/1.73m^2, White patients 71 ml/min/1.73m^2 and Asian patients the highest at 86 ml/min/1.73m^2. Despite this, Black patients were more likely to reach ESRF (43%) compared to Asian (20%) and White (19%) patients, Chi-square $p < 0.0001$, see Fig. 4. White patients have higher complete remission rates at 1 year (19%) compared to Asian patients (6.6%) $p = 0.029$, and Black patients (11%) $p = 0.059$.

Immunosuppression & Remission Status

One hundred patients were treated with immunosuppression and 48 were treated conservatively. The different therapeutic options and usage rates are summarised in Table 3. All treatments were accompanied by steroids in the form of either oral prednisolone or intravenous methylprednisolone. Those receiving immunosuppression were younger, compared to those treated conservatively (median age of 55 vs. 66 yrs. old, $p = 0.0016$). There was no difference in the creatinine, albumin or uPCR at diagnosis between those that were treated either conservatively or immunosuppressed. Additionally, there was no difference in these three parameters at 1 year between these two groups.

Rates of complete remission were highest with CYC at 25%, the least effective immunosuppressants to achieve complete remission were CNIs at 16%. This contrasts to no immunosuppression with a complete remission rate of 6% and AP agents at 24%. The partial remission rate was better with immunosuppression rather than conservative treatment 29% (CYC 38%, CNI 37%, AP 33%). The lowest rate of no remission was in the CYC group at 36% compared to CNI 47%, AP 43% and conservative management 54%. This suggests superiority of cyclophosphamide in our patient cohort, however the results did not reach statistical significance. Black patients were less likely to be treated with CNI (18%) and more likely to be treated with CYC (31%).

Patients undergoing complete remission were younger (mean age 55) compared to both partial and non-responders (61 and 59 respectively). Responders to treatment, irrespective of treatment strategy, had a lower creatinine at presentation; median 87 μmol/L (0.98 mg/dL) compared to 120 μmol/L (1.36 mg/dL) in the non-responders, $p = 0.0116$. Additionally, responders had lower serum albumin at diagnosis compared to the non-responders (albumin 25.5 ± 0.9 gl/L (2.6 g/dL) compared to 28.5 ± 1.2 g/L (2.9 g/dL), $p = 0.0476$), but there was no statistically significant difference in the uPCR.

Table 2 Ethnic diversity, number of patients and percentages of different ethnicities within our cohort

Ethnic group	N (%)
Asian	45 (31)
Black	35 (24)
White	54 (36)
Subgroups	
African	14
Caribbean	21
Bangladeshi	9
Chinese	5
Indian	17
Pakistani	9
Middle Eastern	5

Table 3 Different therapeutic options used in our cohort

Treatment	N (%)
Conservative	48 (32)
Cyclophosphamide	36 (24)
Calcineurin inhibitors	38 (26)
Antiproliferative agents (MMF/azathioprine)	21 (14)
Rituximab/steroid monotherapy/other	1/3/1 (Total 3%)

Discussion

This study corroborates findings from previous studies that conclude patients presenting with impaired renal function are more likely to reach ESRF in MN. [14]. However, at odds with these studies, we have shown that our ESRF patients actually present less nephrotic than the non-ESRF patients. The traditional paradigm is that the worse the proteinuria, the worse the risk of ESRF; patients are even risk stratified for treatment based on the degree of proteinuria in MN [7, 12, 15]. In our cohort, patients who were less nephrotic but with worse renal function progressed to ESRF. This may reflect the reduced glomerular filtration rate attenuating proteinuria, resulting in less severe nephrosis [16].

The other known risk factors for ESRF in MN are gender and age [14]; neither were statistically significant in our group. Further, high lipid levels have been found to contribute to glomerulosclerosis and therefore ESRF independent to the severity of nephrosis [17]; however, we found that patients progressing to ESRF had lower serum cholesterol concentrations, emphasising the attenuated nephrosis in the progression group.

Our data supports the main predictor for progression to ESRF in MN being the serum creatinine at diagnosis. In our cohort, at least, it appears that some of the traditional known risk factors for the development of ESRF with MN are less reliable than previously reported. This is not a trivial matter, as strategies to give toxic treatments for MN are currently based on these risk factors.

A significant limitation to our study is the lack of anti-phospholipase A2 receptor antibody status. This was due to the retrospective nature of the study and the lack of historical serum samples, we are now in a process of collecting anti-PLA2R antibody status of all patients in our tertiary MN clinics. Antibody positivity and titre are important as these are associated with severity and outcome of disease [18, 19].

Ethnic differences

This ethnically diverse group of patients revealed some interesting data. Where details of ethnicity were made available, MN has been reported in homogenous ethnic groups [4, 12, 13]. We found Black patients had worse serum creatinine and lower eGFR (though the eGFR difference was not statistically significant), were more likely to progress to ESRF and were treated more often with cyclophosphamide. There are no comparable studies or reports in Black patients with MN, however, these findings are similar to studies in other renal diseases. It is known that age, sex, race and body weight affect serum creatinine concentration, some of this difference may be due to higher baseline serum creatinine levels found in Black patients and this explains why eGFR differences were not statistically significant [20]. Black patients with lupus nephritis have deteriorating renal function and reduced survival compared to other ethnic groups [21]. Black patients with an eGFR >60 ml/min have a faster rate of decline in renal function irrespective of their albuminuria compared with White patients [22]. The rate of decline persists despite correction of traditional risk factors such as albuminuria, diabetes and hypertension, which suggests an underlying genetic mechanism [22, 23]. There are no studies of MN outcomes in different ethnic groups, however a recent study reviewed the distribution of glomerulopathies in a Southern Californian population. Overall they had lower rates of Black (18.6%) and Asian (8.8%) patients and a larger proportion of Hispanic patients compared to our cohort [24]. There are some reports that Black patients do not respond to CYC as well as other AP agents [21].

There are differences in socioeconomic and biological factors that may explain the faster rate of decline to ESRF in Black patients. Important proposed mechanisms are an interaction of sociodemographic factors with genetic factors such as lower socioeconomic status, chronic stress, psychosocial factors, environmental pollution and differences in access to health care [22]. It should be noted that, like many other studies, we grouped ethnically discrete groups of patients together. For instance, the Asian group included both Indo-Asian and East-Asian patients and the Black group included African and Caribbean patients; these populations are, of course, genetically diverse.

Knowledge about MN has changed significantly as have treatment strategies [4, 12, 13] since the start of the study period. For future studies of MN patients, antibody status and tissue immunohistochemistry of immune deposits and markers of chronic damage correlates are warranted. Furthermore, genomic data would offer insights into the links between ethnicity, gender and outcomes.

Conclusions

This ethnically diverse cohort does not demonstrate the traditional risk profile associated with development of ESRF. Those responding to treatment have more severe nephrotic syndrome, whereas those reaching ESRF have the worst renal function and lowest proteinuria at diagnosis. There are ethnic differences with Black patients presenting with a trend to lower eGFR and having an increased risk of ESRF. This study highlights the importance of careful consideration of therapeutic options, as current risk modelling cannot accurately predict the risk of ESRF. Further studies are required to elucidate the role of antibodies and risk genes.

Abbreviations

ANOVA: Analysis of variance; AP: Anti-proliferative agents; CNI: Calcineurin inhibitors; CYC: Cyclophosphamide; ESRF: End stage renal failure; GFR: Glomerular filtration rate; IQR: Interquartile range; KDIGO: Kidney disease improving global

outcomes; MN: Membranous nephropathy; RASB: Renin angiotensin system blocker; SD: Standard deviation; uPCR: Urinary protein creatinine ratio; USA: United States of America

Acknowledgements
Not applicable.

Funding
Work on this study was supported in part by grants from the European Union, FP7 (EURenOmics grant agreement 2012-305608).

Availability of data and materials
The datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

Authors' contributions
SG participated in data acquisition, analysis, interpretation and wrote the first draft of the manuscript. Authors JC, RP, SBW, MMY, RK and NA participated in data analysis and interpretation and revision of the manuscript. All authors contributed to reading and approving the manuscript.

Authors' information
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Consent for publication
Not applicable.

Ethics approval and consent to participate
Ethical approval was deemed unnecessary per national regulations (NHS Health Research Authority) given the retrospective, non-research nature of the study. All patient identifiable data was anonymised before analysis by the research team. More information available at: http://www.hra.nhs.uk/resources/research-legislation-and-governance/governance-arrangements-for-research-ethics-committees/ The data was collected as part of audits and quality improvement projects at both sites: Research and development office, Royal Free NHS Hampstead Foundation Trust & Clinical Effectiveness Unit, Barts Health NHS Trust.

Publisher's Note

Author details
[1]UCL Centre for Nephrology, 1st Floor, Room 1.7007, Rowland Hill Street, London NW3 2PF, UK. [2]Renal Unit, Barts Health NHS Trust, Whitechapel, London E1 1BB, UK.

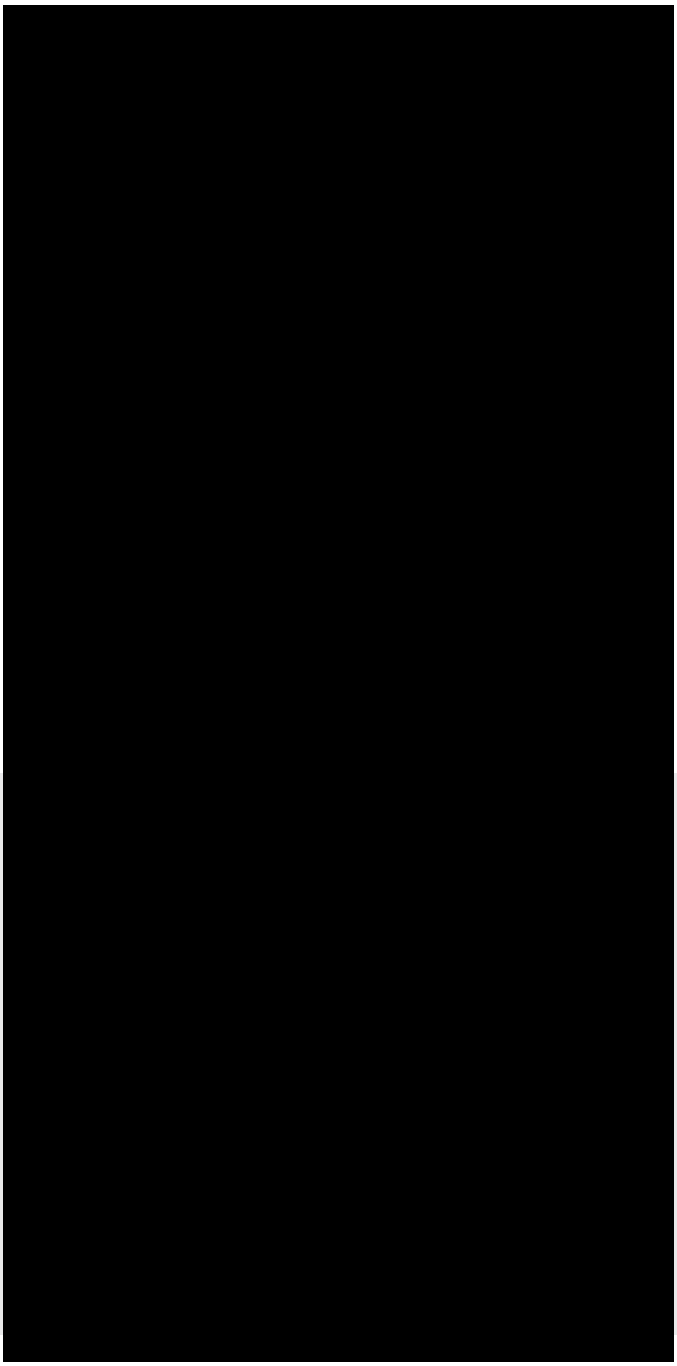
Received: 3 November 2016 Accepted: 8 June 2017
Published online: 21 June 2017

References

1. Fairweather D, Frisancho-Kiss S, Rose NR. Sex Differences in Autoimmune Disease from a Pathological Perspective. Am J Pathol. 2008;173(3):600–9.
2. Rubtsova K, Marrack P, Rubtsov AV. Sexual dimorphism in autoimmunity. J Clin Invest. 2015;125(6):2187–93.
3. Schieppati A, Mosconi L, Perna A, Mecca G, Bertani T, Garattini S, et al. Prognosis of Untreated Patients with Idiopathic Membranous Nephropathy. N Engl J Med. 1993;329(2):85–9.
4. Jha V, Ganguli A, Saha TK, Kohli HS, Sud K, Gupta KL, et al. A Randomized, Controlled Trial of Steroids and Cyclophosphamide in Adults with Nephrotic Syndrome Caused by Idiopathic Membranous Nephropathy. J Am Soc Nephrol. 2007;18(6):1899–904.
5. Ponticelli C, Zucchelli P, Passerini P, Cesana B, Locatelli F, Pasquali S, et al. A 10-year follow-up of a randomized study with methylprednisolone and chlorambucil in membranous nephropathy. Kidney Int. 1995;48(5):1600–4.
6. Hogan SL, Muller KE, Jennette JC, Falk RJ. A review of therapeutic studies of idiopathic membranous glomerulopathy. Am J Kidney Dis Off J Natl Kidney Found. 1995;25(6):862–75.
7. Shiiki H, Saito T, Nishitani Y, Mitarai T, Yorioka N, Yoshimura A, et al. Prognosis and risk factors for idiopathic membranous nephropathy with nephrotic syndrome in Japan. Kidney Int. 2004;65(4):1400–7.
8. Glassock RJ. Diagnosis and natural course of membranous nephropathy. Semin Nephrol. 2003;23(4):324–32.
9. KDIGO Work Group. KDIGO Clinical Practice Guideline for Glomerulonephritis [Internet]. 2012 [cited 2016 Aug 24]. Available from: http://kdigo.org/guidelines/glomerulonephritis-gn/.
10. van den Brand JA, van Dijk PR, Hofstra JM, Wetzels JF. Cancer Risk after Cyclophosphamide Treatment in Idiopathic Membranous Nephropathy. Clin J Am Soc Nephrol. 2014;9(6):1066–73.
11. Cattran DC, Pei Y, Greenwood CM, Ponticelli C, Passerini P, Honkanen E. Validation of a predictive model of idiopathic membranous nephropathy: its clinical and research implications. Kidney Int. 1997;51(3):901–7.
12. Hofstra JM, Branten AJ, Wirtz JJ, Noordzij TC, du Buf-Vereijken PW, Wetzels JF. Early versus late start of immunosuppressive therapy in idiopathic membranous nephropathy: a randomized controlled trial. Nephrol Dial Transplant. 2010;25(1):129–36.
13. Thompson A, Cattran DC, Blank M, Nachman PH. Complete and Partial Remission as Surrogate End Points in Membranous Nephropathy. J Am Soc Nephrol. 2015;26(12):2930–7.
14. Cattran D. Management of Membranous Nephropathy: When and What for Treatment. J Am Soc Nephrol. 2005;16(5):1188–94.
15. Locatelli F, Marcelli D, Comelli M, Alberti D, Graziani G, Buccianti G, et al. Proteinuria and blood pressure as causal components of progression to end-stage renal failure. Nephrol Dial Transplant. 1996;11(3):461–7.
16. Arisz L, Donker AJM, Brentjens JRH, van der Hem GK. The Effect of Indomethacin on Proteinuria and Kidney Function in the Nephrotic Syndrome. Acta Med Scand. 1976;199(1-6):121–6.
17. Kees-Folts D, Diamond JR. Relationship between Hyperlipidemia, Lipid Mediators, and Progressive Glomerulosclerosis in the Nephrotic Syndrome. Am J Nephrol. 1993;13(5):365–75.
18. Hofstra JM, Beck LH, Beck DM, Wetzels JF, Salant DJ. Anti-Phospholipase A2 Receptor Antibodies Correlate with Clinical Status in Idiopathic Membranous Nephropathy. Clin J Am Soc Nephrol. 2011;6(6):1286–91.
19. Hoxha E, Harendza S, Pinnschmidt HO, Tomas NM, Helmchen U, Panzer U, et al. Spontaneous remission of proteinuria is a frequent event in phospholipase A2 receptor antibody negative patients with membranous nephropathy. Nephrol Dial Transplant. 2015;30(11):1862–9.
20. Inker LA, Levey AS. Pro: Estimating GFR using the chronic kidney disease epidemiology collaboration (CKD-EPI) 2009 creatinine equation: the time for change is now. Nephrol Dial Transplant. 2013;28(6):1390–6.
21. Hahn BH, McMahon MA, Wilkinson A, Wallace WD, Daikh DI, FitzGerald JD, et al. American College of Rheumatology guidelines for screening, treatment, and management of lupus nephritis. Arthritis Care Res. 2012;64(6):797–808.
22. Peralta CA, Katz R, DeBoer I, Ix J, Sarnak M, Kramer H, et al. Racial and Ethnic Differences in Kidney Function Decline among Persons without Chronic Kidney Disease. J Am Soc Nephrol. 2011;22(7):1327–34.
23. Stanescu HC, Arcos-Burgos M, Medlar A, Bockenhauer D, Kottgen A, Dragomirescu L, et al. Risk HLA-DQA1 and PLA2R1 Alleles in Idiopathic Membranous Nephropathy. N Engl J Med. 2011;364(7):616–26.
24. Sim JJ, Batech M, Hever A, Harrison TN, Avelar T, Kanter MH, et al. Distribution of Biopsy-Proven Presumed Primary Glomerulonephropathies in 2000–2011 Among a Racially and Ethnically Diverse US Population. Am J Kidney Dis. 2016;68(4):533–44.

P0467 **THE DIRECT USE OF ORAL ANTI-COAGULANTS IN MEMBRANOUS NEPHROPATHY**

Maximilian Wills[1], Sanjana Gupta[1], Alice Gage[1], Neil Ashman[1], Suzanne Forbes[1]
[1]*The Royal London Hospital, United Kingdom*



Strategies for Reduction of Cardiovascular Risk:
Effect of Time and Different Treatments on Lipids in Membranous GN

Introduction

Treatment of membranous nephropathy (MN) is traditionally focused on reducing progression of CKD in the long term, and mitigating complications of nephrotic syndrome in the short term. Although patients with MN are at increased risk of cardiovascular events when nephrotic, the course of hyperlipidaemia and effects of treatment upon this are poorly studied. We conducted a retrospective study to define differences in lipids according to treatments received for MN.

Methods

We identified patients with MN for whom demographic and treatment information was available who had serum cholesterol measurements available at the time of diagnosis, and 3, 6 and 12 months after starting treatment. Differences in serum cholesterol measurements and treatment groups were assessed by two factor ANOVA and Tukey's post-hoc testing using the R statistical computing language.

Results

A total of 234 patients were included in the analysis. 32% of patients were female and the median age at diagnosis was 51 (interquartile range 21 years).

41% of patients were treated with a calcineurin inhibitor (CnI), predominantly Tacrolimus, 23.5% with Cyclophosphamide, 26.4% with supportive treatment (e.g. ACE inhibitor or angiotensin II receptor blocker) and 9% with Rituximab.

Distributions of cholesterol measurements by treatment type and length of follow up are shown in the left panel of the figure. The duration of follow up had a significant effect on total serum cholesterol (P-value 8.6×10^{-9} by two factor ANOVA), with cholesterol falling with increasing time from starting treatment (assessed by Tukey's post-hoc testing).

Serum cholesterol was also significantly different according to treatment used (P-value = 2.9×10^{-7} by two factor ANOVA). The difference in means and associated confidence levels for all possible treatment pairs are shown in the right panel of the figure. A negative difference implies the second listed drug is associated with a lower cholesterol measurement. Cyclophosphamide treatment was associated with lower cholesterol compared to CnI, rituximab and supportive treatment (P-values 2.1×10^{-6}, 0.018, 5.8×10^{-6} respectively). All other treatment comparisons (those in the right panel whose confidence intervals span zero) had similar means.

Conclusion

This preliminary, retrospective analysis of a large cohort of patients with primary MN suggests successful treatment of nephrotic syndrome improves cholesterol, thus modifying the burden of cardiovascular risk. In addition, serum cholesterol may be differentially affected by the treatment used, a factor that ought to be considered in the timing and choice of therapy in the era of less toxic agents.

It is likely that the effects of rituximab may be incorrectly estimated in these data due to its relatively low use (9% of patients). Additionally, time to starting immunosuppressive treatment, time to remission, relevant co-morbidities, and use of lipid lowering agents, are likely to affect treatment response of lipids in MN.