Investigation of the molecular genetic contribution to idiopathic nephrotic syndrome using high-throughput sequencing

Bugarin Diz, Carmen

Awarding institution:
King's College London

Investigation of the molecular genetic contribution to idiopathic nephrotic syndrome using high-throughput sequencing

Carmen Bugarin Diz
1542754

Department of Medical and Molecular Genetics
July 2021

Thesis submitted to King's College London in fulfilment of the degree of Doctor of Philosophy

Declaration

I hereby declare that the work presented in this PhD thesis is my own.

Acknowledgements

First and foremost, I would like to thank my supervisor Dr. Ania Koziell for her endless support and expertise during the last four years and giving me the opportunity to pursue a PhD. I would also like to thank my second supervisor Prof. Michael Simpson for taking me on in his group and providing crucial knowledge and advice to make this project possible.

I am grateful for all the members of the Simpson group for creating a friendly and motivational environment. Particularly thanks to Ines and Jake for welcoming me since day one and for their friendships over the past years. Thanks to Nick for his advice and useful chats about life or coding and for taking the time to read some of my work. I am no less indebted to Teresa for her contagious positive energy, especially during the final months. I would also like to thank Dr. Richard Dillon and all the CanGen lab for their support and help.

Finally, I would like to thank my family for their unconditional support. My mum for always being there for me and my dad for all his encouragement. Both have made great sacrifices to get me here. My auntie Ines to put up with me when work became stressful. Also, big thanks to my second family here in the UK, the McFarlane clan. Most of all I would like to thank my partner, Jenny, for everything.

Abstract

Nephrotic syndrome (NS) is a rare kidney disease resulting from malfunction of the primary ultra-filtration unit in the kidney, the renal glomerulus, leading to excessive leak of protein into urine. NS has an annual incidence of 2 and 7 per 100,000 children and adults, and a prevalence of 1 to 15 per 100,000 depending on the ethnicity; NS is more common in African and South Asian populations. However, despite its rare disease status, it remains one of the most common kidney diseases to affect children and adults. It has a devastating impact on the health of affected individuals, with around 20% of cases developing end stage kidney failure and 60% of the severe group experiencing disease recurrence post kidney transplant.

Mendelian inheritance appears only to explain around 30% of cases. Inheritance may be autosomal recessive or dominant with variable penetrance and more recently X-linked has also been described. To date, causal genetic variants have been identified in 67 genes in NS patients. However, the molecular genetic mechanisms underlying the remaining 70% remain poorly understood and are likely to fall into a complex genetic category.

The aim of this study was to identify causal genetic variation of NS, focusing on the 70% of cases currently unexplained by single mutations in previously established nephrotic syndrome genes. A cohort of 485 deeply phenotyped patients was available for analysis; all have undergone whole exome sequencing or whole genome sequencing. Data was analysed by applying computational approaches including linkage and association analyses. Based on this, a link with HLA was identified, confirming that despite a lack of typical inflammatory markers, NS in both children and adults falls into the category of an autoimmune disease.

Table of Contents

Table of Figures

Table of Tables

Chapter 1 – General Introduction

1.1 Renal system

1.1.1 Kidney structure and function

The main function of the renal system is to filter blood to remove waste products and excreted them into urine. The kidneys are complex organs constituted by many specialised cells and are located in the retroperitoneal space. These organs can be divided into two main regions: cortex and medulla. The structural and functional unit of the kidney is the nephron, and an adult human kidney contains approximately one million of these (1). The nephrons are responsible for the filtering of the blood, waste removal and reabsorption of needed substances. Each nephron contains a network of capillaries called glomerulus where the filtering of the blood takes place and a tubule, that concentrates substances secreted into the urine and recovers any important solutes from the primary urine to the blood. The glomeruli and most of the proximal tubules are found in the kidney cortex whereas the medulla contains the majority of the distal portions of the tubules. Thus, in the medulla is where the concertation of urine is performed (1) (Figure 1).

The kidney function is crucial for a healthy fluid composition in the body, correcting any variations that might occur due to food intake, metabolism, environmental factors and exercise. Thus, kidneys are responsible for the excretion of metabolic end products or foreign products like toxins and drugs. Additionally they also produce important enzymes and hormones such us renin, erythropoietin and 1,25 di-hydroxy vitamin D3 (2).

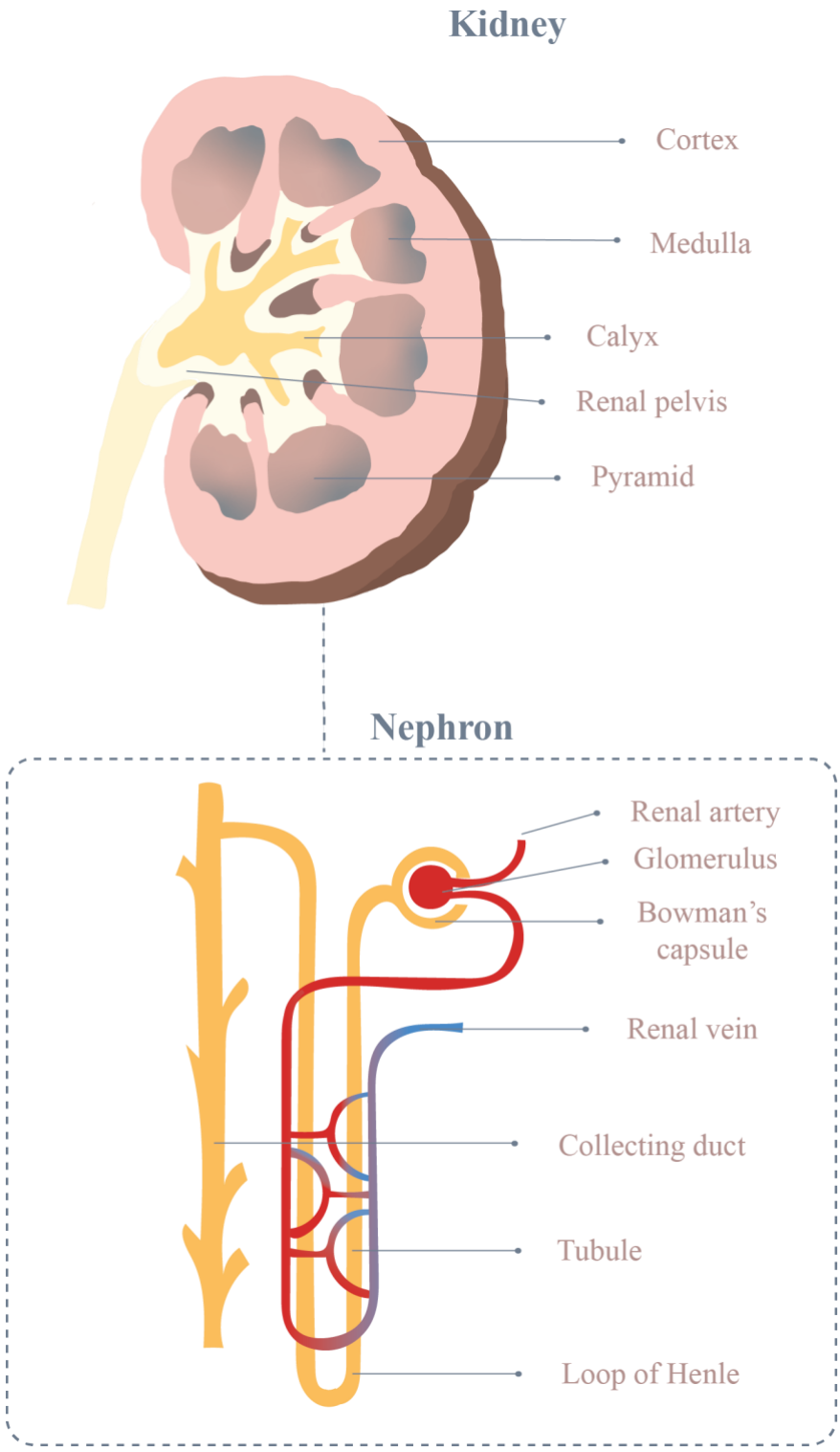


Figure 1. Morphology of the kidney and schematic representation of the nephron. The kidney is divided into two main regions: cortex and medulla. The nephron is the structural and functional unit of the kidney

The process of urine formation begins with the blood entering the kidney through the renal artery into small arterioles that form a capillary tuft in the glomeruli. These glomerular capillaries are supported by basement membrane of highly specialised epithelial cells known as the glomerular filtration barrier (GFB), where waste products with a low molecular weight are filtered and an ultrafiltrate of plasma is formed. The filtered fluid is collected in the Bowman's capsule and enter the renal tubule. Most of the glomerular filtrate is reabsorbed at the loop of Henle into the renal vein specially water and ions, although some additional substances are secreted. Then, the urine enters the distal convoluted tubule and finally the collecting duct, where it is transported through the kidney medulla to empty at the renal pelvis (Figure 1). The extracellular fluid in this region of the kidney has a much higher solute concentration than plasma and the main function of the medullary structures is the concentration of urine. The final product enters the renal pelvis where is transported to the bladder and is finally excreted from the body.

1.1.2 The glomerular filtration barrier

The glomerular filtration barrier is the first filtering unit in the kidney located in the Bowman's capsule in the renal cortex. The GFB is constituted by fenestrated endothelial cells, the glomerular basement membrane (GBM) and glomerular epithelial cells known as podocytes (Figure 2). Endothelial cells and podocytes share a protective negative charge net made by glycocalyx contributing to the permeability properties of the barrier (3). Therefore, the GFB is a highly sophisticated macromolecular sieve with size and charge restricting characteristics. Molecules with a molecular weight above 15-20 kDa are normally unable to traverse the barrier (4).

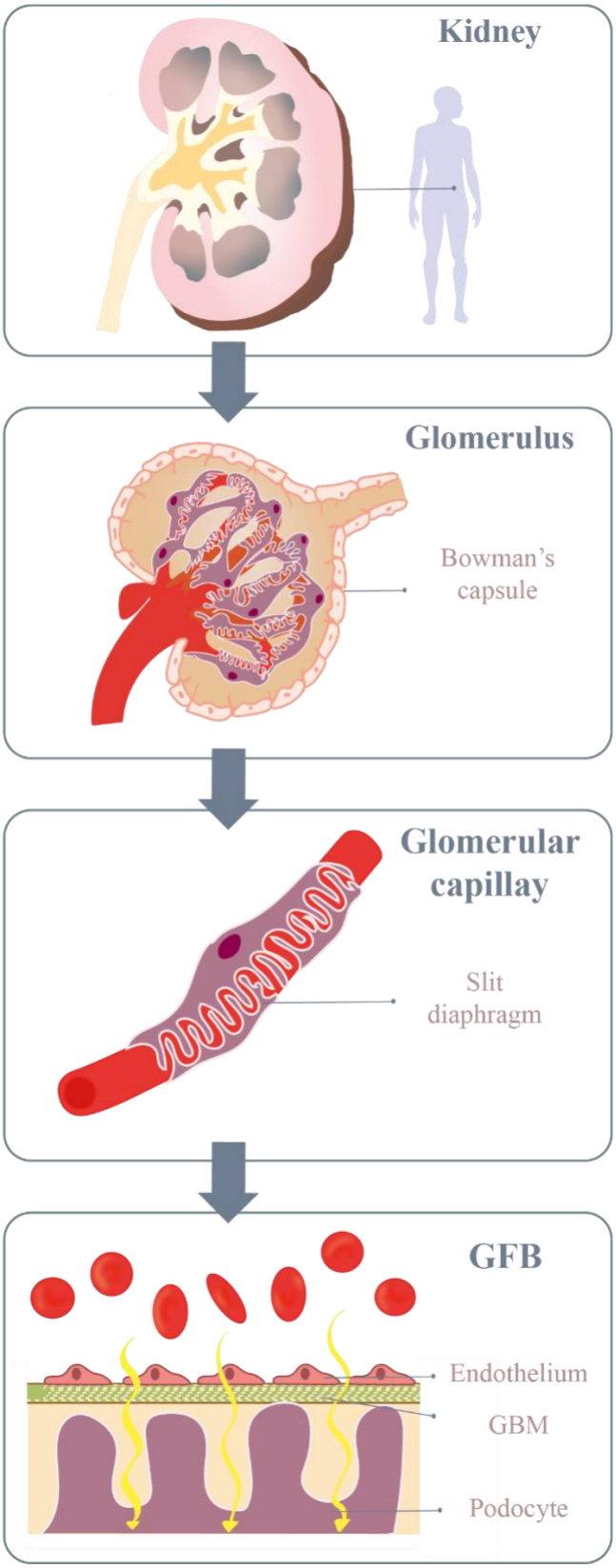


Figure 2. Schematic view of the glomerular filtration system. There are approximately 1 million of glomeruli in the cortex of each kidney. The glomerulus is formed by multiple capillaries. Each capillary is constituted of fenestrated endothelium and surrounded by the glomerular basement membrane and podocytes.

The intercellular junctions form by podocytes which are called slit diaphragm, leave space between the cells to allow the ultrafiltration of molecules. The glomerular basement membrane restricts the passage of large molecules, with the podocytes and slit diaphragms acting as the framework to keep it in compression (5). The attachment of the glomerular basement membrane and podocytes is made via foot processes which act as a filter contributing to size limit of the GFB, as well as reinforcing it against the relatively high tensile forces resulting from the capillary pressure of 40mmHg. Additionally, podocytes are thought to have a contractile function to confer elasticity to the GFB. (6). Thus, the integrity of the actin cytoskeleton that keeps the podocyte architecture is key for the function of the glomerular filtration barrier.

1.2 Nephrotic syndrome

1.2.1 Clinical definition and presentation

Nephrotic syndrome (NS) is a rare kidney disease, with an estimated annual incidence of 2 to 7 cases per 100,000 children and adults and prevalence of 1 to 15 cases per 100,000 individuals depending on ethnicity and region (7). Epidemiological studies in the United Kingdom (UK), United States (US) and Canada have reported a higher incidence of nephrotic syndrome in Africans and South Asians compared to Europeans (8, 9). NS may occur as an isolated kidney defect known as idiopathic nephrotic syndrome (INS) or be part of a syndrome involving other organ systems such as diabetes mellitus, systematic lupus erythematosus or myeloma and lymphoma, where the glomerulus becomes affected as a secondary hit (10, 11). Mitochondrial diseases have also been linked to some of the renal manifestations present in nephrotic syndrome (12).

NS manifests clinically with various physiological changes related to abnormal fluid balance. It is defined by a triad of symptoms such as excessive leakage of protein into the urine (proteinuria), swelling of the body (oedema) and also low levels of plasma albumin (hypalbuminaemia). Mostly NS symptomatology is a sequence of connected events: proteinuria leads to low levels of plasma albumin and a decrease of the intravascular oncotic pressure resulting in sodium retention and movement of water into the interstitial space to cause nephrotic oedema. The loss of other key plasma components into urine leads to metabolic disturbances like dyslipidaemia, a pro-thrombotic tendency, anaemia and low vitamin D (13).

A diagnosis of nephrotic syndrome is established by blood and urine investigations and if indicated, a kidney biopsy. Ranges vary but generally, proteinuria is confirmed when the protein excretion is more than ≥3.5 g/day and the protein-to-creatinine ratio > 2000 mg/g. Urinary protein loss is a prognostic biomarker and is considered an independent risk factor for cardiovascular morbidity and mortality. Hypoalbuminemia is confirmed when serum albumin <30 g/L. Additionally, some patients present severe hyperlipidaemia with elevated cholesterol cholesterols >10 mmol/L. Although renal biopsy in patients with NS might be useful for treatment and prognosis especially in adult disease, it is not performed on most paediatric patients who are initially responsive to treatment, as it has been found to be of limited benefit (14, 15). Current guidelines advocate renal biopsy in nephrotic syndrome individuals that do not respond to treatment at all ages, and if atypical features are present such as haematuria, hypertension or extra-renal features either compatible with a syndrome or multisystem disease which may be suggestive of a secondary NS (16, 17). Renal biopsy remains an integral part of the management of patients with adult-onset nephrotic syndrome in view of the higher incidence of secondary causes such as membranous nephropathy

and other pathologies such as IgA nephropathy, and in differentiating atypical presentations of proteinuria in childhood masquerading as NS.

1.1.2 Pathogenesis

The molecular pathogenesis of INS results from malfunction of the first filtering unit in the kidney, the glomerular filtration barrier. As previously mentioned, the GFB is a highly sophisticated macromolecular sieve with size and charge restricting characteristics (18). It allows free flow of water and small solutes but restricts the transit of molecules >15 kDa. In NS, there is a disruption of the permselective properties of the GFB allowing the loss of important proteins leading to further disruption of body homeostasis (19). While damage to any of the three components of the GFB, fenestrated endothelium, extracellular glomerular basement membrane or podocytes, results in kidney disease, podocytes appear to be the main cells targeted by injury and have a central role in the initiation and progression of glomerular disease (20). Podocyte damage or loss is hard to restore due to the limited capacity of cell regeneration. However, the causes leading to podocyte disruption remain unclear in INS as well as the underlying mechanisms that allow the recovery of the podocyte function using immunosuppression treatment (19).

The efficacy of different immunosuppressive treatments including steroids in many patients indicates the importance of the immune system in the pathogenesis of INS. Abnormal function of T lymphocytes had been shown in some patients with an increase of $CD8^+$ cells and a decrease of $CD4^+$ circulating T cells, during active phases of disease (7, 21). B cells are also another potential immune candidate as treatment of INS patients targeting CD-20 antigen on the surface of B cells, significantly reduces the number of relapses (22). Moreover, disease onset and relapses are frequently

triggered by infectious diseases, mainly viral, or allergic episodes suggesting that the activation of the immune system plays an important role in INS (21).

An unknown "circulating factor" has also been hypothesised to be involved in the INS pathogenesis by increasing the permeability of the GFB to albumin. Evidence of this phenomena have emerged after some patients experience disease recurrence post kidney transplantation from a healthy donor (23). Additionally, clinical and in-vivo studies have shown that injection of plasma from INS patients induces proteinuria (24). Recent data on the efficacy of a new technique for treatment of individuals with nephrotic syndrome recurrence post renal transplant, Liposorber LDL apheresis, supports that this may be a lipid related molecule (25).

1.1.3 Histology

Examination of renal biopsies from patients with INS have identified different histological appearances: minimal change disease (MCD), focal segmental glomerulosclerosis (FSGS), membranous nephropathy, membranoproliferative glomerulonephritis (MPGN) and complement 3 glomerulopathy (C3G) (13) (Figure 3). MCD is seen in the majority of nephrotic syndrome diagnosis that response to treatment especially in childhood. Here the glomeruli appear normal under light microscopy but loss and fusion of podocyte foot processes are evident on electron microscopy. In contrast, FSGS is more frequently but not exclusively associated with affected individuals that response poorly to steroids. The pathological changes that take place develop focally (not in all glomeruli) and segmentally (only in parts of a glomerulus) and are visible with light microscopy (26). Membranous nephropathy is less common in nephrotic syndrome cases and very rare in children. It is characterised by immune complex deposits located between podocytes and the GBM (27).

Moreover, there are other histological variants of INS such as tubular dilatation or diffuse mesangial sclerosis (DMS), but they are not frequent (28).

Historically, the first clinical classification of INS was based on histological findings aligned with outcome and response to treatment. MCD was thought to be associated with good prognosis and response to treatment, whereas FSGS was associated with the worst outcome and resistant to treatment (29). Whilst this is very broadly true in simple nephrotic disease, the pathology may be more complex with recent studies providing evidence supporting a common aetiology for MCD and FSGS, suggesting they are different manifestations of the same progressive disease. Thus, patients with MCD can develop FSGS over time (30, 31). Overall, response to injury is influenced by multiple factors such as type of disease, treatment or genetic background, with some cases progressing to renal scarring, whereas others manage the initial localised insult and progress to repair (32).

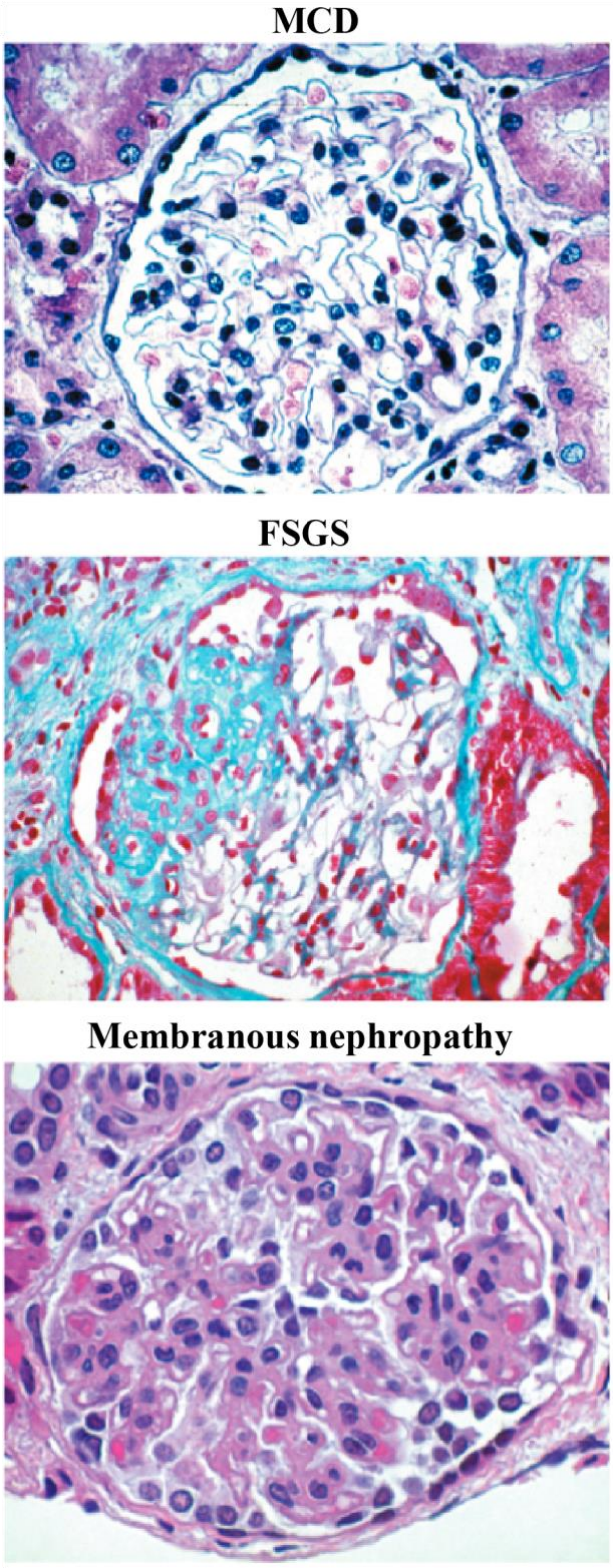


Figure 3. Renal biopsies from patients with INS. Histopathology slides from children with MCD, FSGS and membranous nephropathy. Image taken from Eddy AA et al (13).

1.1.4 Treatment

At present, only non-directed treatment strategies centred on generalised immunosuppression are available. These alleviate symptoms and minimise complications but rarely result in a 'cure' (33). The standard initial treatment for INS is corticosteroid therapy, normally given as oral prednisone for at least 4 - 6 weeks. The time from onset to remission is considered an important prognostic factor for the disease in the sense that response in children should be seen within this time (34). In adults, technically 4 months is considered the cut off for classifying the patient as steroid resistant. Patients that experience frequent relapses or steroid dependency require steroid-sparing agents such as levamisole, mycophenolate mofetil (MMF) and tacrolimus, individually or in combination (33). According to UK NICE guidelines, if these steroid sparing agents are ineffective, patients are treated with a chimeric monoclonal antibody called rituximab that targets CD-20 antigen on the surface of B cells depleting the activation of these cells (35). This is often successful in arresting disease progression. For patients who fail to respond and are classed as resistant to conventional therapy, low-density lipoprotein (LDL) apheresis using the Liposorber system may be offered (25). Management of INS varies depending on patient's condition and side effects, drug efficacy, clinician's preference and drug availability by country. However, all of these drugs have the main objective to achieve long term nephrotic remission by either decreasing the activation of immune cells (36) or in some instances potentially by direct action on podocytes (37).

Long term outcomes for INS patients vary depending on response to treatment. Around 5 – 10 % patients achieve complete remission after initial treatment with prednisolone alone. However, the vast majority will require further courses of prednisolone and/or a second-line agent to reach remission. Patients who experience

frequent relapses and develop secondary treatment resistance, and do not respond to rituximab, have a future risk of end-stage renal disease (ESRD) and require renal replacement therapy (38, 39). If this leads to kidney transplant, cases have a 60% risk of disease recurrence in the graft, especially if initially, some treatment response was observed (40).

1.2 Clinical classification of nephrotic syndrome

Response to steroid medication has led to an arbitrary classification of INS: steroid sensitive nephrotic syndrome (SSNS) or steroid resistant nephrotic syndrome (SRNS). However, many patients have an initial response to treatment (around 50% in childhood) but then unpredictably, develop subsequent resistance to steroids, known as secondary SRNS, for reasons that are currently unclear, and these may progress to total treatment resistance. Thus, it is increasingly recognised that the divide is often artificial and not specifically indicative of the underlying pathology (Figure 4).

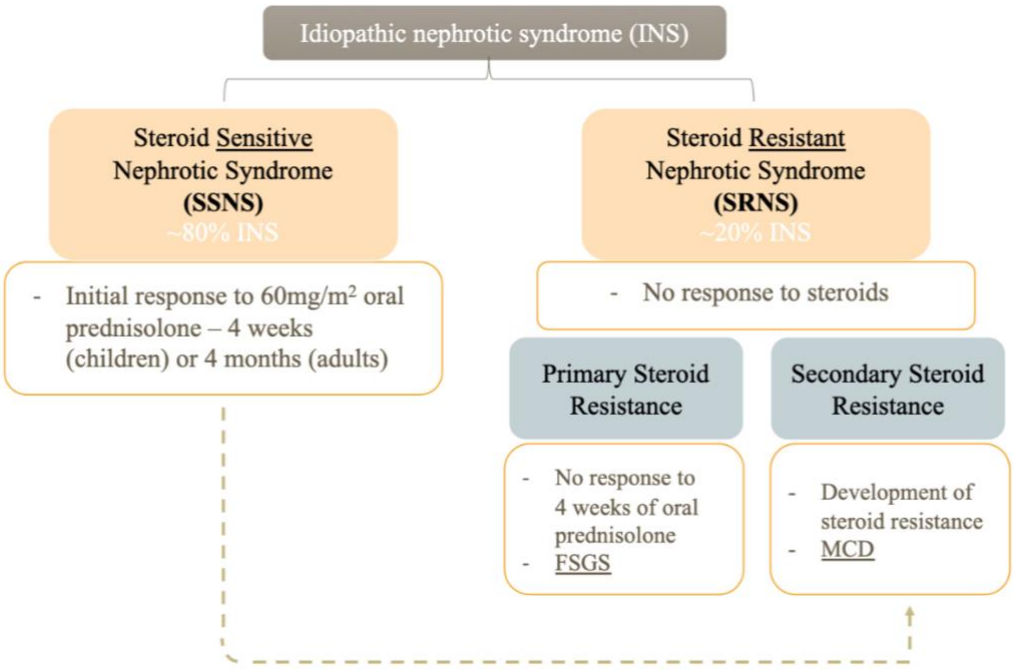


Figure 4. Stratification of idiopathic nephrotic syndrome. Based on response to treatment INS can be classified into two groups: SSNS and SRNS. Some patients do not response to treatment also known as primary SRNS whereas others do response initially but later develop a resistance to steroids knowns as secondary SRNS

1.2.1 Steroid sensitive nephrotic syndrome (SSNS)

SSNS is responsible for 80-90% of total cases of the disease and responds to immunosuppressive agents like corticosteroids and calcineurin inhibitors. The histological subtype most common is MCD (41). Although the majority of patients achieve remission after treatment, clinical course varies with different relapse rates often accompanied by dependence on steroid administration (42). At least half of the patients experience frequent relapses that can lead to steroid dependency and complications, such as increased morbidity and poor quality of life. However, less than 5% of SSNS patients progress to end-stage renal disease (43).

1.2.2 Steroid resistance nephrotic syndrome (SRNS)

Around 10-20% of INS cases do not respond to steroid within 4 weeks in childhood and 4 months in adults, and as such are classed as steroid resistant (SRNS). Primary

SRNS is defined as non-response to prednisolone within 4 -6 weeks in children and 4 months in adults, whereas secondary SRNS occurs when responsiveness to prednisolone or other steroids is lost at some point in the therapeutic pathway. Those cases that do not respond to steroids may be multidrug resistant and also not response to other immunosuppressant agents (44). However, some cases may still respond to drugs, suggesting that the steroid resistant phenotype is not an absolute. Primary SRNS constitutes the third most prevalent cause of ESRD in the first two decades of life (45). The most common histological subtype is FSGS, although patients may have MCD at disease onset and before developing FSGS lesions. Rare forms of inherited nephrotic syndromes such as diffuse mesangial sclerosis (DMS) and Finnish Type congenital nephrotic syndrome may be found in early life. Additionally, approximately 33% of SRNS patients experience disease recurrence post kidney transplant (40).

1.3 Genetics basis of idiopathic nephrotic syndrome

INS is a heterogeneous disease, as genetic and environmental factors do play a role in modulating phenotype. The identification of disease-causing mutations has improved our understanding of the pathogenesis of INS. As phenotypes can vary, conventional clinical assessment and diagnostic techniques may not provide complete answers. As such, genetic testing can be crucial in the clinical setting as an adjunct to diagnosis and subsequent management of patients with nephrotic syndrome (19). Some gene mutations may also stratify cases and have prognostic value helping identify guidelines on treatment selection.

1.3.1 Genetics of SSNS

While specific loci are known or suspected to increase the risk of SSNS, only one single gene defect has been described as a cause of Mendelian forms of SSNS, *EMP2*, which was detected as causative in autosomal recessive inheritance (46, 47). However, other studies in familial SSNS did not find any link with *EMP2* or other genes that could be responsible for monogenic forms of the disease despite previous described associations (48). A genetic predisposition has been suspected since different epidemiologic studies have reported multiple families with different generations affected by SSNS. Additionally, there are families with a combination of affected members that are steroid sensitive and steroid resistant (46, 49). Thus, patients might share a common genetic factor that predisposes to SSNS but in response to yet unidentified environmental triggers, could develop different reaction to treatment. A study of 59 SSNS families identified variants associated with disease in the human leukocyte antigen (HLA) region, precisely in the *HLA-DQ* gene, implying the importance of the adaptive immunity in the molecular mechanisms of SSNS (47).

The HLA region, also known as the major histocompatibility complex (MHC), is the most polymorphic region in the entire genome and it has been associated with the greatest number of human diseases, including immune-mediated renal diseases (50). The HLA locus contains a cluster of genes that are crucial for the immune system function. It is divided into three subclasses of genes: class I; *HLA-A,-B* and *-C*, class II; *HLA-DR,-DQ,* and *-DP* and class III; *TNF, HSP70* and *C2-C4* (51). In addition to being a gene-rich and extremely polymorphic region with considerable variation across different populations, it has an extensive linkage disequilibrium (LD) (52). This results in multiple genetic markers being coinheritance because they are nonrandomly associated due to their close proximity. Therefore, these genomic characteristics make

the accurate identification of the exact alleles responsible for disease signals very challenging.

HLA class II genes have long been associated with SSNS, although the molecular mechanisms of how these genes are involved in the pathogenesis of the disease remain unknown (47). HLA class II molecules are only expressed by antigen-presenting cells (APCs), such as dendritic cells, mononuclear phagocytes and B cells. Their function is to present self or foreign peptides to $CD4^+$ T cells, helping the adaptive immune system to send an appropriate response to infection while maintaining immune tolerance to self-antigens (Figure 5) (50, 51). Thus, it has been hypothesised that antigen presentation might be in some way impaired in SSNS patients (53).

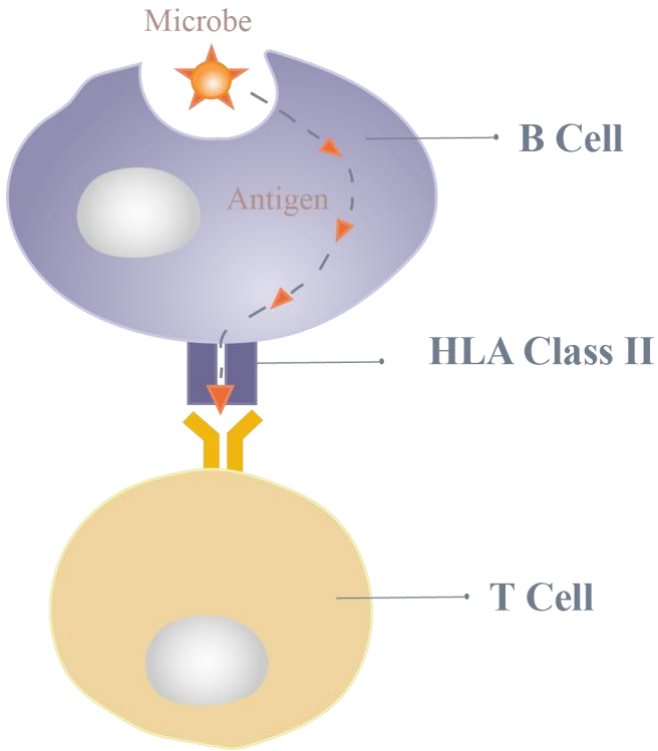


Figure 5. Antigen presentation by MHC class II. HLA class II molecules are only expressed by antigen presenting cells (APCs) such as B lymphocytes, dendritic cells and macrophages. A fragment of a foreign peptide is phagocytosed inside the B cell and transported to the cell surface to be presented to T cells.

1.3.1.1 Risk alleles

Genome-wide association studies (GWAS) have also been used to shed light into the aetiology and pathogenesis of SSNS. The increasing abundance of good quality patient genomic data available has facilitated the discovery of combinations of common genetic variants that are known to increase the risk of having rare and complex disorders. This approach uses genotype data from a cohort of individuals that share same phenotype to identify alleles that are present in greater or lower frequency compared to an unaffected ethnically matched control population. To date, at least three independent GWAS studies have reported an association with HLA-DR/DQ region and SSNS in children and adults of European and also Japanese ethnicity. Specifically, using HLA imputation the studies described risk and protective alleles in *HLA-DQA1* and *HLA-DQB1*. These loci explained 6-10% of the genetic risk for SSNS (54-56).

1.3.2 Genetics of SRNS

Advances in high throughput sequencing technologies have resulted in sequencing cost declining making possible the screening of patients at first presentation with SRNS for a monogenic cause of disease. As a result, targeted panel, whole exome (WES) and whole genome sequencing (WGS) have become widely used for the identification of disease associated genes in SRNS (57). Genetic diagnostics provides important information to better stratified SRNS, improving treatment and transplant management. Studies of Mendelian forms of SRNS suggest patients do not response to immunosuppressive agents and are less likely to present recurrent disease after transplantation compared to those without a known monogenic cause (58).

1.3.2.1 Mendelian disease

The majority of genes responsible for monogenic forms of SRNS were identified by studying family pedigrees through linkage analysis. When sequencing data from large family pedigrees are available, genetic variants can be studied for segregation with disease status. This approach narrows the search space for the causative variant, as it is expected to segregate with the phenotype status within the family, therefore variants present in unaffected members can be discarded (59). Additionally, consanguineous families have been particularly effective in increasing the statistical power to identify novel or known pathogenic variants using homozygosity mapping. Monogenic SRNS can be autosomal recessive (AR) or autosomal dominant (AD), with variable penetrance and more recently X-linked has also been described (60, 61). Typically, early onset of the disease tends to be caused by mutations in kidney developmental genes, resulting in malformation of GFB, generally under an autosomal recessive model with high penetrance. In contrast, later onset of the disease more frequently affects genes responsible for the regulation of the actin cytoskeleton in the podocytes, under autosomal dominant inheritance with incomplete penetrance (62). Here the phenotype may be more variable through incomplete penetrance.

The discovery of the first mutations in genes reported to be causal for Mendelian SRNS was crucial in understanding the importance of the podocyte dysfunction in the pathogenesis of the disease. Various genetic mutations were shown to cause podocyte abnormalities. Often studies use terminology loosely when describing established SRNS genes as “podocyte genes” even when these are not specific or highly expressed within this cell type. Furthermore, not only the podocyte function is disrupted in

SRNS, but also the signalling crosstalk mechanisms with other glomerular cell types, potentially influenced by epigenetic events and unknown pathogenic events (63) (Figure 6).

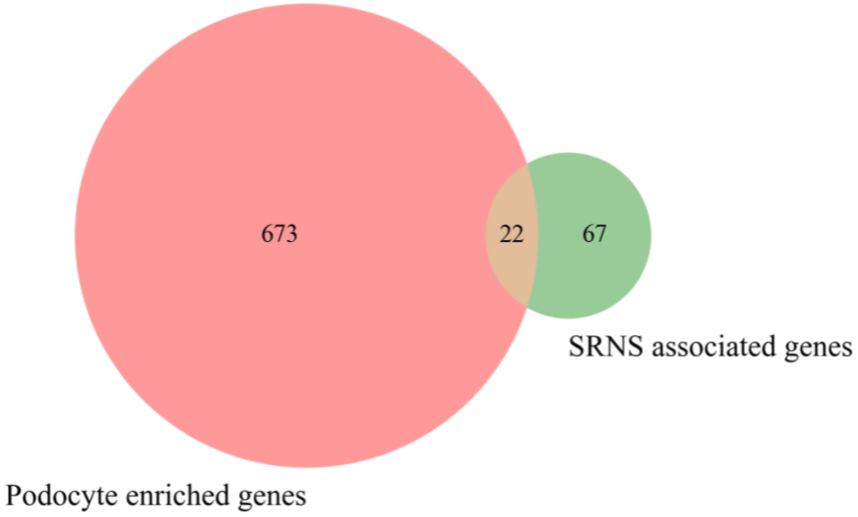


Figure 6. Venn diagram of gene overlap between podocyte enriched genes and SRNS associated genes. The pink circle represents 673 podocyte enriched genes and the green circle 67 genes directly associated with SRNS. There is an overlap of 22 genes between the two groups shown in yellow.The list of genes in which it expression was enriched in human podocytes was extracted from single cell RNA-Seq data from the studied made by Gillies CE et al (64).

The *WT1* gene, was originally identified in 1990 as a candidate gene for Wilms' tumour through studies of patients with 11p13 deletions causing WAGR syndrome (Wilms' tumour, anirida and genitourinary abnormalities) (65). *WT1* encodes an ubiquitously expressed transcription factor with a regulating domain (exons 1-6) and a DNA binding domain (exon 7 – 10). Beyond being a tumour suppressor gene, *WT1* has a key role in the control of genitourinary development. Nonetheless, it was not until 1991 when mutations in *WT1* were detected mostly within exon 9 under autosomal dominant inheritance in patients with Denys-Drash syndrome (DDS), which is characterised by SRNS, intersex and Wilms' tumour (66, 67). With further molecular genetic analysis the spectrum of *WT1* mutations was expanded to a related

and equally rare condition that also presents with SRNS, Frasier Syndrome, a triad of SRNS, intersex and gonadoblastoma (68, 69). The examination of related phenotypes from apparently different disorders detected an overlap at a molecular level since cases of familial Frasier syndrome in one generation followed by Denys Drash in the next (70, 71).

NPHS1 and *NPHS2* were subsequently cloned in 1998 and 2000 respectively and found to cause autosomal recessive congenital nephrotic syndrome (CNS) and autosomal recessive early onset SRNS (72, 73). *NPHS1* mutations cause massive proteinuria by resulting in a developmental defect of the GFB. Whilst proteinuria is detectable from birth, the oedema may not appear until 6 weeks of age once glomeruli start to mature. *NPHS1* encodes nephrin, a transmembrane protein member of the Ig-superfamily of adhesion molecules. It has complex mechanisms of action including participation in multiple protein-protein interactions, a role in autophagy, cellular signalling and is a key component of the slit diaphragm junctions located between podocytes within the GFB. *NPHS1* mutations result in disruption of protein and malformation of slit diaphragms in the kidney. Through a founder effect, NPHS1 mutations are causative in 98% of Finnish children with this syndrome (CNS Finnish Type I), although non-Finnish cases have a lower incidence and are genetically more heterogeneous (74). *NPHS2* is the only established SRNS gene exclusively expressed in the podocytes and encodes a membrane protein called podocin. Mutations in this gene have been associated with SRNS under an autosomal recessive model of inheritance which may be familial or rarely sporadic through de novo mutations (75). Dysregulation of podocin protein function leads to alterations in the slit diaphragm architecture (73). Podocin also has multiple functions such as protein-protein

interactions including with the intracellular domains of nephrin helping to recruit to lipid rafts, important signalling platforms within the podocyte membrane (76).

Another gene that was discovered to be mutated in patients with early onset nephrotic syndrome *LAMB2*. Mutations in *LAMB2* are detected in patients with Pierson's syndrome, characterised by early onset SRNS that rapidly progresses to ESRD and ocular abnormalities (77, 78). Laminins are one of the most predominant components in the glomerular basement membrane and they play an important role in cell motility and adhesion (79)

Together, mutations in *NPSH1, NPHS2, WT1* and *LAMB2* are responsible for a 66.3% of congenital nephrotic syndrome cases at first year of age (80, 81). However, from the very first causal gene discoveries in SRNS primarily based on traditional positional cloning techniques, advances in next generation sequencing technology over the last 20 years including whole exome and whole genome sequencing have dramatically enabled new disease gene discovery, at times in the absence of large and interbred pedigrees. Consequently, the list of genes causing nephrotic syndrome has dramatically increased. Currently, mutations in at least 67 genes have been reported to cause SRNS (Table 1). Despite the genetic heterogeneity, mutations in *NPHS1* and *NPHS2* are responsible for approximately 44% of the SRNS cases reported in the literature partly through reporting bias in childhood disease (Figure 7). Whilst identifying the prevalence of mutations in rare diseases is problematic, a lack of replication in wider cohorts of SRNS raises the question whether some more recent SRNS genes specifically those identified in highly interbred families, may actually be segregating with the family rather than disease especially in the absence of confirmatory functional data.

Overall, only 30% of patients with family inherited or early onset have a monogenic mutation in one of the established SRNS genes documented in the literature (82). Despite the possibility that some mutations have been missed, or exist in genes that have not yet been discovered, this supports an emerging hypothesis that many SRNS cases may not follow Mendelian patterns of inheritance especially in sporadic cases with an absent family history.

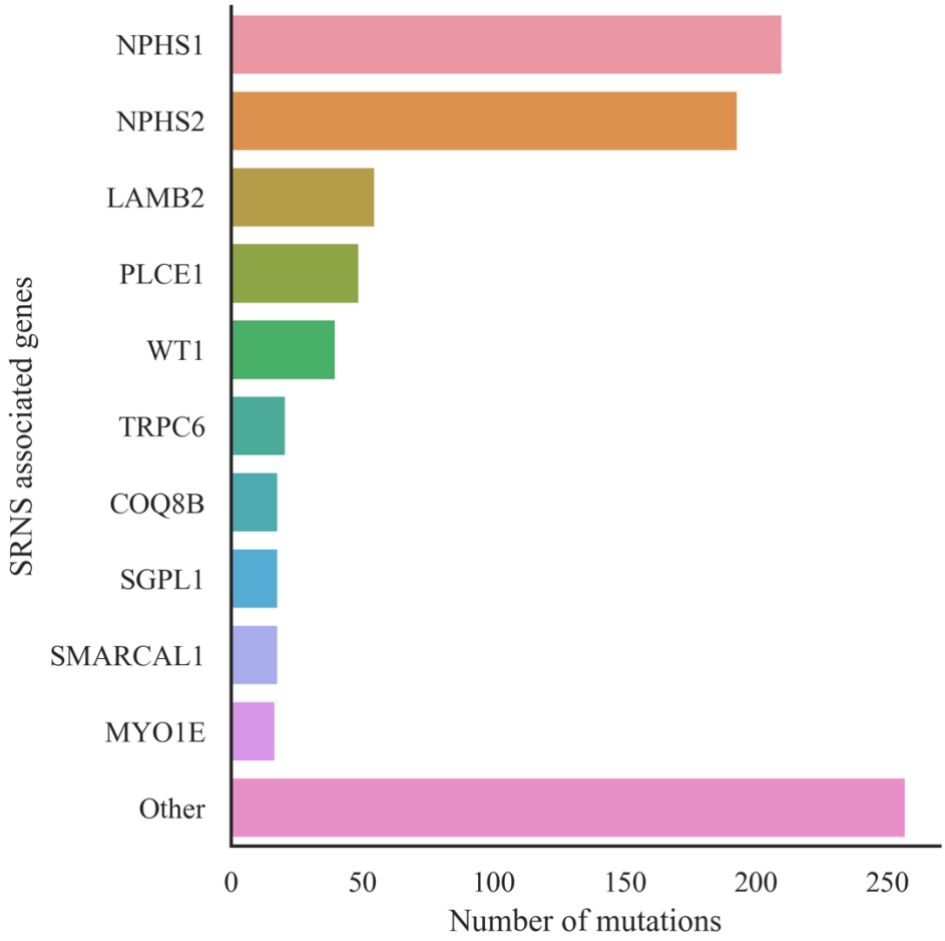


Figure 7. The ten most commonly mutated genes in SRNS patients. A total of 896 mutations have been reported for the phenotype 'Nephrotic syndrome, steroid resistant' in HGMD 2020.4 release (83). The breakdown of the ten genes with the highest number of mutations is showned in the barplot. The 'other' label represents the sum all of the mutations found in 57 genes known to cause SRNS.

Table 1. List of 67 genes considered directly associated with nephrotic syndrome.

Gene	Model	Disorder
Slit diaphragm associated and adaptor proteins		
NPHS1	AR	CNS/NS
NPHS2	AR	CNS/NS
PLCE1	AR	Early onset NS
KIRREL1	AR	NS
CD2AP	AR	Early onset NS, HIV nephropathy
FAT1	AR	NS, Hematuria, Tubular ectasia
Actin cytoskeleton components		
ACTN4	AD	Adult onset NS
INF2	AD	NS, Charcot-Marie-Tooth
MYH9	AD	Adult onset NS
MYO1E	AR	FSGS
ARHGDIA	AR	CNS/NS
ARHGAP24	AD	Adult onset NS
ANLN	AD	Adult onset NS
MAGI2	AR	NS
PODXL	AD	NS
SYNPO	AD	FSGS
KANK1	AR	Early onset NS
KANK2	AR	Early onset NS
KANK4	AR	Early onset NS
NPHP1	AR	Early onset NS
ANKFY1	AR	NS, CKD
GSN	AD	NS, Amyloidosis
GAPVD1	AR	NS, CKD
ITSN2	AR	NS
DLC1	AR	NS
TBC1D8B	AR	NS
AVIL	AR	NS
NEIL1	AR	NS
Glomerular basement membrane proteins		
LAMB2	AR	Pierson syndrome, CNS with ocular abnormalities
ITGB4	AR	NS
ITGA3	AR	Epidermolysis bullosa and NS
CD44	AR	Pretibial bullous skin lesions, NS
COL4A1	AD/AR	NS
COL4A3	AD/AR	Alport disease
COL4A4	AD/AR	Alport disease
COL4A5	X-linked recessive	Alport disease
FN1	AD	Glomerulopathy
LAMA5	AR	NS

Apical membrane proteins		
TRPC6	AD	Adult onset NS
EMP2	AR	Early onset NS
Nuclear proteins		
WT1	AD	Denys-Drash Syndrome, Early onset NS
LMX1B	AR	NS, Nail-Patella Syndrome
SMARCAL1	AR	Schimke immuno-osseous dysplasia
PAX2	AD	Adult onset NS
LMNA	AD	NS, Familial partial lipodystrophy
NXF5	X-linked recessive	NS, Heart block disorder
NUP85	AR	NS, FSGS
NUP93	AR	NS, FSGS
NUP107	AR	NS, FSGS
NUP133	AR	NS, FSGS
NUP160	AR	NS, FSGS
NUP205	AR	NS, FSGS
XPO5	AR	Early onset NS
Mitochondrial proteins		
COQ2	AR	Mitochondrial disease, Nephropathy
COQ6	AR	NS, Deafness
PDSS2	AR	NS, Leigh syndrome
COQ8B	AR	CNS
Other intracellular proteins		
APOL1	AR	Adult onset NS
PTPRO	AR	Early onset NS
CRB2	AR	Early onset NS
DGKE	AR	NS, Hemolytic-uremic syndrome
ALG1	AR	Congenital disorder of glycosylation
CUBN	AR	NS, Epilepsy
TTC21B	AR	NS with tubulointerstitial involvement
WDR73	AR	Galloway-Mowat Syndrome
SGPL1	AR	NS
LCAT	AR	NS

1.3.2.2 Risk alleles

In 2010, a genome-wide association study identified two risk alleles in *APOL1* that were significantly associated with FSGS and chronic kidney disease in the African American population (84). This study was possible because of genomic resources from the 1000 Genomes Project (1KGP) (85). The international efforts performed by the 1KGP established a catalogue of human variation from diverse ancestry allowing the study of common and rare variants in the African population. Findings were then replicated in other studies demonstrating the two variants account for some of the excess risk of kidney disease in Africans compared to Europeans (86).

Both risk alleles are coding variants that lead to amino acid changes that alter the function of *APOL1*. They are normally described as: G1 allele, two non-synonymous single nucleotide polymorphisms (rs73885319, rs60910145) and G2 allele, in-frame deletion of two amino acid residues (rs71785313). Inheritance of two risk alleles leads to an increased risk of FSGS and CKD, whereas inheritance of one risk allele causes a much smaller risk. Thus, both alleles have a recessive model of inheritance (84). Additionally, a much smaller effect is found in patients with heterozygous G1 compared with heterozygous G2 (87). The allele frequencies of G1 and G2 in the African population are 0.22 and 0.13 respectively. Approximately 13% of African Americans have any of the two *APOL1* risk alleles (G1+G1 or G2+G2) causing a 7-to-30-fold increased risk of renal disease (88). These variants are common and also have large effects on disease susceptibility.

1.4 Generation of human genetic variation profiles using high-throughput sequencing

The next milestone in human genetics after the publication of the first draft sequence of the human genome in 2001 (89), was the development of massive parallel sequencing technologies known as “high-throughput sequencing” in 2005. This was enabled by two ground-breaking studies that revolutionised the field of genetics by describing a method that allowed accurate sequencing of the entire bacterial genome (90, 91). This technique based on PCR allows the amplification and sequencing of multiple DNA fragments simultaneously, increasing the speed and reducing the overall cost. Before which, Sanger sequencing was the methodology that dominated the genome sequencing field, where only a single DNA fragment was processed at a time.

Affordable genome-scale sequencing has revolutionised the medical sector increasing the understanding of the molecular genetic basis of multiple diseases, especially Mendelian disorders (92). This technology allows the study of human genetic variation where genomes from patients can be compared with healthy individuals in order to identify disease causing mutations. Thus, the catalogue of genetic variants underlying human diseases, in monogenic and complex conditions, has changed medical decisions and treatments, making possible a personalised medicine (92-94) (Figure 8).

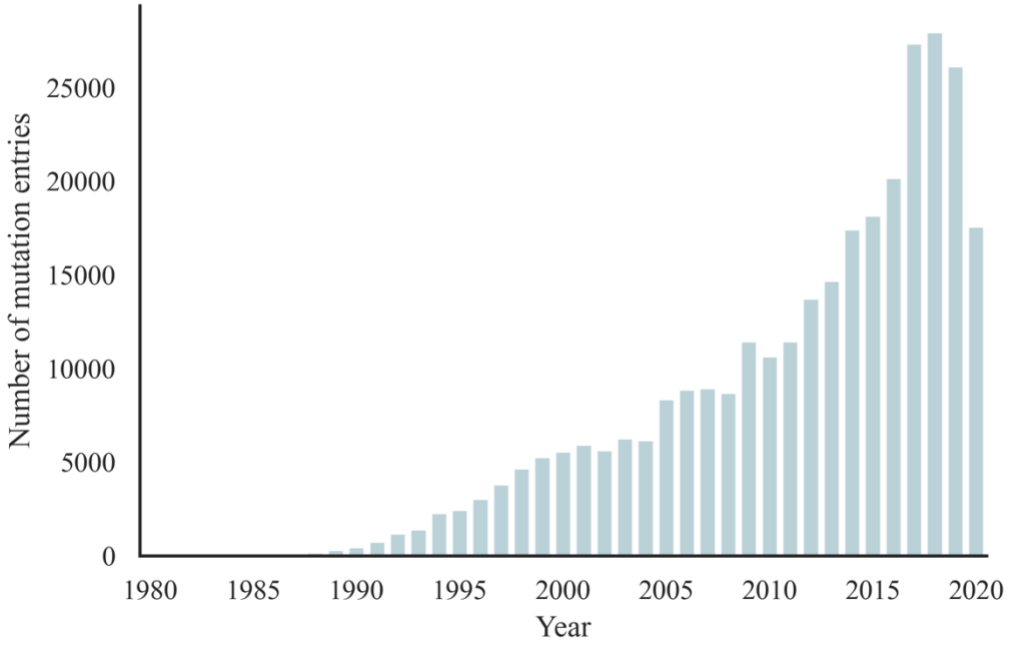


Figure 8. Number of disease-associated mutations by year of publication. The cumulative number of mutations reported in the Human Gene Mutation database (HGMD) each year from 1980 to 2020. Following the development of high-throughput technologies the number of disease-associated mutations has increased dramatically, effectively doubling since 2010 (Data source obtained from the HGMD website).

High-throughput sequencing encompasses a number of different methodologies that share the following steps: library preparation, cluster generation, sequencing and data analysis. Each technique uses specific protocols to transform raw data into meaningful information, and the output varies depending on the platform used (95). Illumina sequencing platform has been the most successful and has generated more than 90% of the world's sequencing data (96). Sequencing technologies can be divided into two main categories depending on read length where data output can have long or short reads (97). Some regions of the genome are highly complex with multiple long repetitive elements that are undetectable using short read strategies. Thus, long-reads are preferred to map these complexity regions in a single continuous read. However, short reads are cheaper and less prone to error.

1.4.1 Library preparation of whole exome and whole genome sequencing

High-throughput sequencing can be applied to entire genome known as whole genome sequencing (WGS) or to specific genomic regions that encode proteins known as whole exome sequencing (WES). The human genome contains 180,000 exons and they constitute just 1% of the total (~30 Mb) (98). The exomes encompass all the annotated protein coding genes that number approximately ~22,000 (99). Restricting the size of the genomic material studied enables the sequencing of more individuals at a deeper depth and in a lower cost. Majority of alleles that are known to underlie Mendelian disorders lie on the exon regions of the genome, altering the protein folding and other cellular processes (Figure 8). Furthermore, the output data generated by WES is easier to store and quicker to analyse and process (100). Despite being a cost-effective option, the analysis of exomes also presents limitations as it does not assess the impact of non-coding variants, especially regulatory regions. Ascertainment of disease causing single-nucleotide (SNV) and structural variants (SV) is less reliable in WES compared with WGS (101, 102).

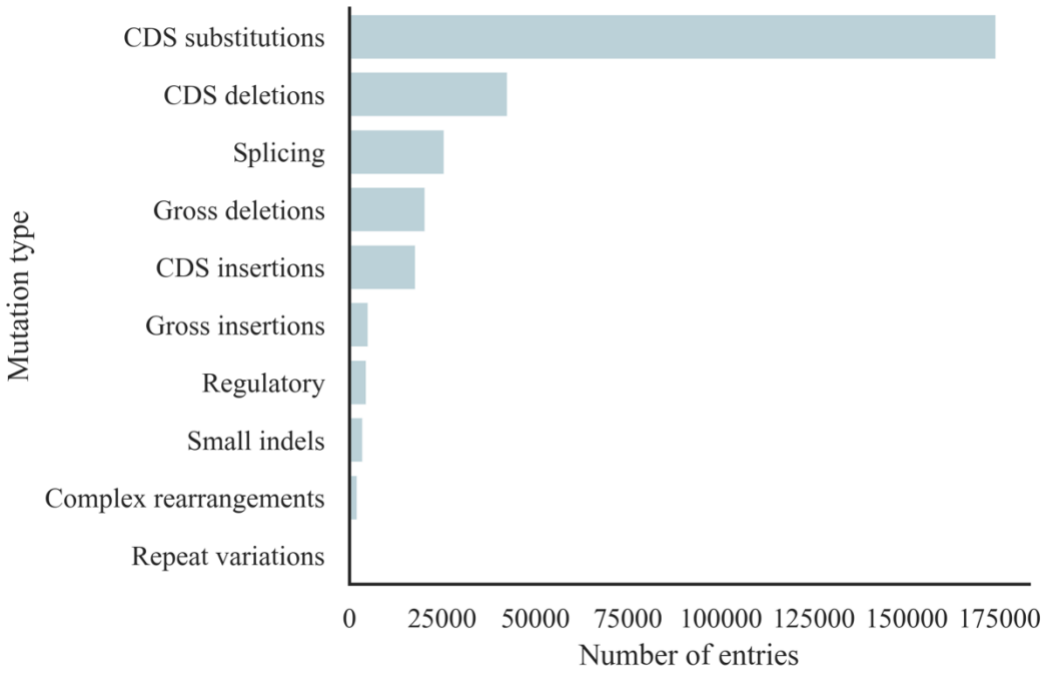


Figure 9. Number of entries by mutation type in the HGMD database. HGMD does not include either somatic or mitochondrial mutations. More than 90% of mutations reported are located in the coding regions also known as coding sequence (CDS).

Over the last ten years, exome sequencing and low-coverage whole-genome sequencing have been used to study the human genetic variation within genome profiles. A typical individual's genome differs from the reference human genome at approximately 4.1 to 5.0 million sites (103). The majority of the variants are single nucleotide changes and short indels that are frequently found in the population of the sample. The total number of sites that differ from the reference genome varies dramatically depending on the population ethnicity (104). Profiles from African ancestry contain greater variation with the highest number of non-reference sites, consistent with the demographic history of human origins (105).

The first step in the sequencing process for WES and WGS is the library preparation, required to generate the template that will be use for the sequencing reaction (106) (Figure 9). Genomic DNA is extracted in laboratory from blood samples and checked

for high quality. Once purified, the DNA is randomly cut into small fragments of similar length (~150 bp) by sonification leaving the double strand with uneven ends where one strand has a few base pairs more than the other. In order to repair the uneven ends, a single adenine base is added to form an overhang also known as A-tailing reaction. Subsequently, the DNA fragments are linked to specific sequences of a couple of base pairs long called adapters (Figure 10). These adapters are crucial to start the DNA reading reaction in downstream analyses but also for the identification of samples, as they contain a specific barcode sequence for each sample. When whole exome sequencing is conducted an extra step is required in order to capture the coding regions of the genome. An in-solution capture method developed by Agilent was used on the samples studied in this dissertation (107). This technology uses a pool of custom RNA oligos known as probes to selectively hybridise to exons in the genome (Figure 10). These probes are biotinylated and bind to special magnetic streptavidin beads in the solution, allowing to wash away the parts of genome that are not exons. Lastly, the overall template of DNA fragments generated by the genomic library is attached to a solid surface containing multiple flow cells where billions of sequencing reactions take place simultaneously (108).

1.4.2 Sequencing of short DNA fragments using Illumina platform

Illumina is the most widely used platform for short-read sequencing technology with a wide range of protocols where the read length goes up to 300 bp. Illumina dye sequencing is achieved by solid-phase bridge amplification and a ‘sequencing by synthesis’ approach (95) (Figure 9). The aim is to amplify the DNA fragment generating hundreds of identical strands of DNA.

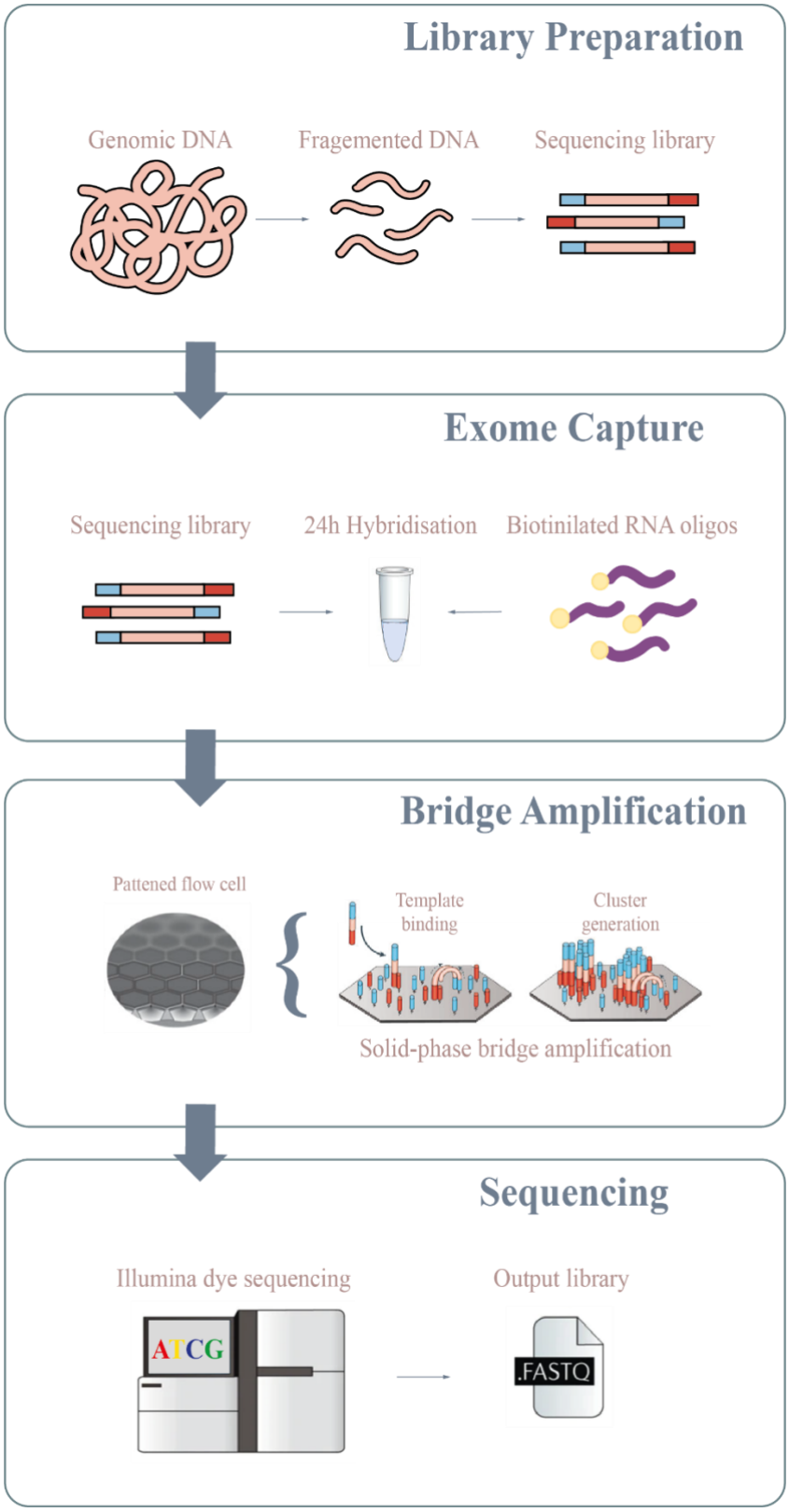


Figure 10. Generation of human genetic profiles. Diagram showing the most important steps of whole exome sequencing: library preparation, exome capture, bridge amplification and sequence by synthesis. For whole genome sequencing the same steps are followed except for the exome capture. Some of the images were adapted from Goodwin S et al and Hardwick SA et al (95, 109)

After the library preparation, the ligated products are denatured and attached to a flow-cell surface. Each fragment is then amplified into distinct clonal clusters via bridge amplification (110). Multiple cycles of annealing, extension and denaturation happen resulting in clonal amplification of all fragments all over the flow cell simultaneously. Once the clonal amplification is complete, the templates are ready for sequencing. The DNA polymerase adds a fluorescently tagged dNTP to the DNA strand at a time. After each round, the single base incorporated in the DNA fragment is detected by a sequencing machine that reads the fluorescent signal as the four bases (adenine, thymine, guanine, cytosine) have an unique emission. The result is highly accurate for base by base sequencing (95).

1.4.3 Alignment, variant calling and annotation

The raw data generated by the sequencing process requires a series of processing and quality control steps before embarking on downstream genetic analyses. The output data is stored in a FASTQ file (111) (Figure 10), which is a common file format for sharing sequencing data containing the DNA sequence also known as read and quality scores. The PHRED software (112) assigs a quality score for each base estimating the probability of error and is calculated as:

$$Q_{PHRED} = -10\ x\ log_{10}\ (P_{incorrect\ basce\ call})$$

The bases with a low-quality score (PHRED <20) are trimmed from the reads. The first step is to align the reads from the FASTQ file to a reference genome (112). All samples studied in this dissertation were aligned to GRCh37 (Genome Reference Consortium human build 37). Reads are analysed in order to find fragments with overlapping areas, called contigs, and compared to the reference genome for variant identification. The aligned reads are then stored in Binary Alignment Map (BAM) file

format (113). There are multiple quality control (QC) steps that take place at this stage where the duplicates from the PCR reaction are removed, the GC content and the strand bias is reported to identify potential contaminations and the quality scores are recalculated. Variant calling is then performed by identifying the positions of the positions of the reads that differ from the reference genome. Variants are normally stored in Variant Call Format (VCF) file (114). The number of variants identified through WES and WGS varies depending on many factors such as the technology used, variant calling approach, coverage and ethnicity of the samples. The total number of variants ranges from ~90,000 in the exome regions to 4 million variants in whole genome (85, 115, 116). Finally, some annotation can be added to each variant to facilitate genetic analyses, specifically for disease-associated gene discovery. Often, variants are annotated with population-based allele frequencies for each allele and the prediction of the consequence of the variant at a protein level (missense, splicing…) (117). Additionally, scores for pathogenicity prediction can be added too.

1.5 Identification of disease-causing genetic variation

The interpretation of the thousands of variants that are identified by high-throughput sequencing technologies remains challenging as genes fall along a spectrum of pathogenic and benign variation. While successful gene discovery can lead to a better understanding of the disease and improve the clinical care for patients, the analysis of genomes or exomes without the right evidence framework can lead to false positives (118). Statistic evidence is required in order to identify a causal relationship between a gene and phenotype. Thus, a causal variant must be significantly enriched in cases compared to controls (119).

When the phenotype being studied is a monogenic disorder some assumptions are generally made for the searching of causative mutations. In Mendelian diseases, a single mutation is sufficient to cause the disease. The causal variant must be rare in the general population and is likely to affect the function of the protein encoded by the gene. Additionally, individuals that carry the mutation present the phenotype (also known as complete penetrance) and the same gene must be mutated in other unrelated affected individuals. Therefore, variants with a high allele frequency for the populations with the same ancestry to the patients studied are normally rejected. Coding regions of the genome are prioritised and synonymous variants that are not expected to have an effect on the resulting protein product are discarded (120). Furthermore, the suspected model of inheritance of the disease is also taken into consideration and variants can be filtered on zygosity; if a dominant model is assumed only variants that are heterozygous would be considered.

In 2015 the American College of Medical Genetics (ACMG) published the first guidelines on the clinical interpretation of variants identified in patients (121). Since then, these recommendations have been refined over time and have been adopted internationally by genetic diagnostic laboratories for rare diseases and familial cancers across different countries (122). The guidelines recommend using a five-tier system where variants are classified as pathogenic, likely pathogenic, of uncertain significance, likely benign or benign. The term “likely” corresponds to 90% certainty of either benign or pathogenic classification (121). Moreover, likely pathogenic and pathogenic are considered to be evidence that can be used in health setting for clinical decision making. The classification is based on evidence from different sources such as minor allele frequency (MAF), segregation through family studies and computational predictions. However, adherence to these standards and guidelines is

not compulsory in research settings and should not be considered inclusive or exclusive to the use of other procedures.

1.6 Aims and overview

Idiopathic nephrotic syndrome is a rare disorder of the kidney glomerulus. Proteinuria, the hallmark of this disease, occurs as a consequence of the loss of the normal permselective properties of the glomerular filtration barrier (GFB) with the primary target for cellular injury being the podocyte layer. At least 67 genes have been associated with affected individuals demonstrating classical Mendelian inheritance. There is also considerable developmental and functional genomic evidence that other mechanisms including common genetic variation play a key role in glomerular filtration both in health and disease. As such, the heterogeneous nature of disease make the study of its molecular genetics and clinical subtypes challenging. Furthermore, the rarity of INS, specifically SRNS forms, has dramatically limited the size of patient cohorts available for research purposes. Therefore, the majority of the SRNS studies are inadequately powered.

In this dissertation I studied, to my knowledge, one of the biggest cohorts of SRNS patients in Europe with stringent phenotypic data available to improve our understanding of the molecular genetic basis of the disease. A comprehensive description of the cohort with phenotype and clinical features is shown in chapter 3. Additionally, this project contains three distinct objectives in which whole exome and whole genome sequencing were used in combination with multiple study designs and analytical strategies. All methodology used can be found in chapter 2. The aims can be summarised as:

1. To evaluate rare genetic variation in the coding regions of established SRNS genes in this cohort. This was achieved in chapter 4 by WES-based screening of all genes that are currently known to cause nephrotic syndrome across all samples. The individuals that were whole genome sequenced were transformed to WES.
2. To identify rare variants that are potentially pathogenic in previously undescribed associated nephrotic syndrome genes. A stringent variant filtering pipeline is described in chapter 5, combined with pedigree segregation and burden test analysis to identify potential causal variants.
3. To study the contribution of common genetic variation within the entire genome in SRNS patients. In chapter 6, I described the analysis that led to the identification common genetic variants that influence the risk of SRNS by genome-wide association study.

Overall, the main goal of this project was to investigate the contribution of rare and common variants to the pathogenesis of INS, specifically SRNS. Lastly, in chapter 7, I explained the limitations of the study and the major conclusions made from these analyses. Furthermore, I discussed the impact of some of these results for patients and to the current knowledge of nephrotic syndrome pathogenesis.

Chapter 2 – Materials and Methods

2.1 Study participants

2.1.1 Cases

The study cohort included a total of 422 individuals (including sporadic cases and probands, excluding duplicates and affected/unaffected family members) with familial nephrotic syndrome as well as sporadic cases. All had a diagnosis of either primary or secondary steroid resistant nephrotic syndrome (SRNS) and were classified based on standard clinical criteria and renal biopsy. All individuals underwent rigorous deep phenotyping to ensure phenotypic accuracy and consistency. This type of standardisation was needed to avoid variation in clinical criteria and standards of clinical recording, which can make the study of relatively small rare disease cohorts challenging. Any misclassified cases of nephrotic syndrome such as IgA nephropathy were removed from further study.

Assessment included direct patient contact and data collected from hospital records, the Renal Rare Disease Registry (RaDaR) and the UK Renal Registry (UKRR). 16 patients were subsequently re-classified as steroid sensitive (SSNS) by Dr. Ania Koziell but included in the study in view of severe disease and relatively short follow up. In total, there were 14 families (4 duos, 6 trios, 2 families with five members, 1 family with six members and 1 family with twelve members). Genomic data was then gathered through two independent sequencing projects, one sequencing the whole exome and the other the whole genome. The diagnostic demographic of the study cohort is illustrated in Figure 11 below.

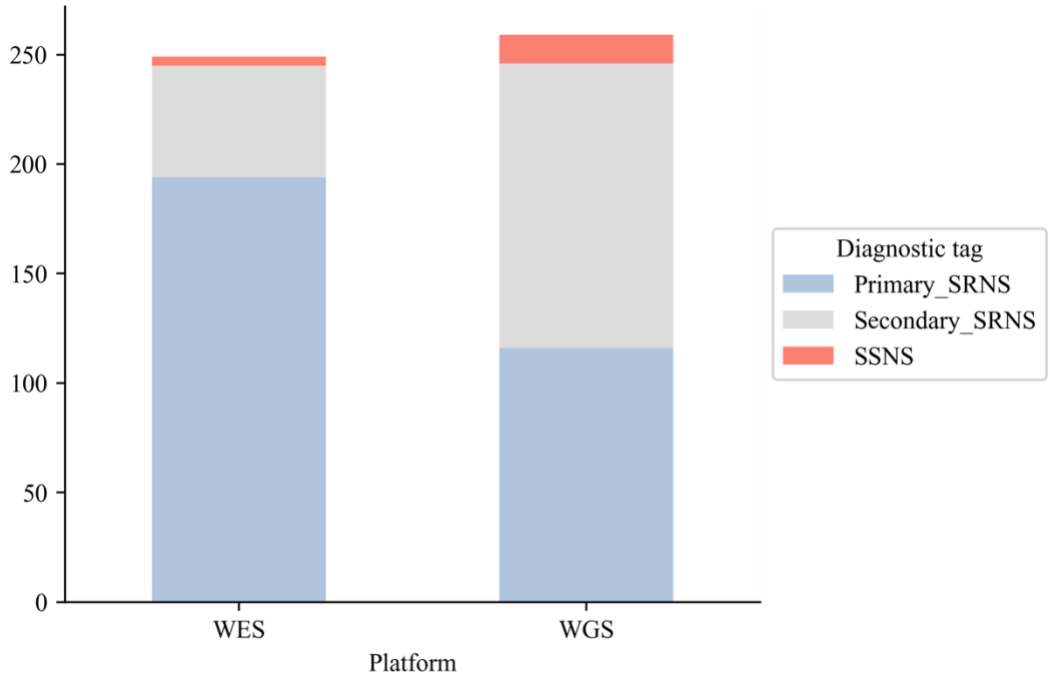


Figure 11. Breakdown of nephrotic syndrome subtypes by platforms. Three nephrotic syndrome diagnostic tags and their counts sequenced in two platforms (WES and WGS) as part of two distinct research projects. Primary SRNS is shown in blue, secondary SRNS in grey and SSNS in red. For WES, majority of samples were diagnosed with primary SRNS whereas for WGS more than half of the samples were diagnosed with secondary SRNS. The number of samples diagnosed with SSNS was very low.

(i) KCL-GSTT BRC SRNS Programme (WES)*:* 256 patients with primary nephrotic syndrome were recruited nationally, around half through the NephroS/RaDaR projects (https://renal.org/rare-renal/patient/nephrotic-syndrome-0) and the remainder through the specialist glomerular disease clinics at Evelina London Children's Hospital and Guy's hospital. Samples were whole exome sequenced by the Genomics Core at KCL/Guy's and St Thomas Hospital Trust (GSTT) Biomedical Research Centre. Sequencing data was aligned and annotated using an in-house pipeline (123). Families were categorised using letters.

(ii) National Institute for Health Research (NIHR) BioResource for the 100,000 Genomes Project Rare Diseases Pilot (WGS)*:* A further 277 cases were recruited as part of another national rare disease study where 13,037 participants, of whom 9,802

had a rare disease as well as some of their close relatives were enrolled from 57 National Health Service (NHS) hospitals in the United Kingdom and 26 hospitals in other countries as part of the Cambridge NIHR BioResource Rare Disease whole genome sequencing study, a pilot for the Genomics England 100k project. There were originally 18 study domains comprising of rare diseases and healthy controls (Table 2). The SRNS domain recruited 277 deeply phenotyped individuals, including families and sporadic cases. SRNS cases were recruited nationally from different centres around the UK, with around 70% from Evelina London and Guys hospital. Other centres were again Bristol Children's, Cardiff and Vale University Health Board, Newcastle Children's Hospital, North Bristol NHS Trust, Nottingham University Hospitals NHS Trust, Queen Elizabeth Hospital Birmingham, Southampton Children's Hospital, Colchester General Hospital and Manchester PCT. Sequencing data was aligned and annotated using the pipeline at the University of Cambridge High Performance Computing Service (HPC). Families from this cohort were categorised using numbers.

2.1.2 Controls

For whole exome sequencing, 1000 control individuals from the 1958 British Birth Cohort were analysed using the same in-house pipeline (123). The raw data (FASTQ files) was aligned and annotated in the same way as the cases. This is a nationally representative resource that includes individuals born during a specific week in 1958 (124). This cohort was chosen because is an unselected population and the use of controls that are large groups of cases of specific diseases has potential to create situations that identify spurious associations.

For whole genome sequencing, in the absence of healthy control samples, individuals were selected across 11 rare disease domains and 2 domains (The UK Biobank Extreme Red Cell Traits and Technical Controls) with apparently healthy individuals. Cohorts with other kidney phenotypes (membranoproliferative glomerulonephritis [PMG]), cancer (multiple primary malignant tumours [MPMT]) and large effect associations reported for common variants (such as pulmonary arterial hypertension [PAH] and primary immunodeficiency diseases [PID]) were excluded. All domains were sequenced by the same Illumina platform and processed in the pipeline at the University of Cambridge High Performance Computing Service (HPC) as part of the pilot study for the 100,000 Genomes Project. (102) (Table 2).

Table 2. NIHR BioResource study domains. A total of 18 study domains part of the pilot study for the 100,000 Genomes Project with their number of individuals. For the control selection only 13 domains were used.

NIHR BioResource Domains	Total
100,000 Genomes Project - Rare diseases pilot (GEL)	4889
Pulmonary Arterial Hypertension (PAH)	1216
Primary Immunodeficiency Diseases (PID)	1430
Bleeding and Platelet Disorsders (BPD)	1206
Extreme Red Cell Traits (UK Bionbank)	766
Inherited Retinal Dystrophy (IRD)	736
Neurological and Developmental Disorders (NDD)	688
Multiple Primary Malignant Tumours (MPMT)	633
Intrahepatic Cholestasis of Pregnancy (ICP)	306
Steroid Resistant Nephrotic Syndrome (SRNS)	277
Hypertrophic Cardiomyopathy (HCM)	269
Stem Cell and Myeloid Disorders (SMD)	267
Cerebral Small Vessel Disease (CSVD)	260
Neuropathic Pain Disorder (NPD)	210
Membranoproliferative Glomerulonephritis (PMG)	195
Tenchnical Controls (CNTRL)	73
Leber Hereditary Optic Neuropathy (LHON)	72
Ehlers-Danlos Syndrome (EDS)	23

Additionally, different control populations were utilised at several stages throughout this study. For the association tests, allele frequencies from the relevant ethnic groups were extracted from different online databases explained in the section 2.5.4 Variant Annotation.

2.2 Ethical approval

Work preceding the analyses in this thesis included writing study designs, application for ethical approval and data capture. To obtain the clinical data presented in this work, ethical protocols 13/EE/0325 (NIHR BioResource – Rare Diseases Cambridge, BRIDGE) and 09/H0106/80 (Radar/NephroS: NURTuRE study) were approved, and informed consent obtained under the supervision and guidance of Dr Ania Koziell. All patients provided informed written consent for the use of their DNA and clinical data.

2.3 DNA extraction and storage

DNA was extracted from peripheral blood with the Gentra Puregene Blood Kit (Qiagen). The extraction procedure was performed by Clinical Genomics Lab hub at Guy's Hospital. Aliquots of DNA were stored for future use in case confirmation within a clinically accredited lab was required for the purposes of genetic diagnosis.

2.4 Sequencing

2.4.1 Sample preparation

Samples selected for whole exome sequencing were assessed for quality by quantifying their concentration with a Qubit Fluorometer (Invitrogen) using the BR and HR dsDNA assay kits, according to the manufacturer's instructions.

2.4.2 Whole exome sequencing

For WES samples, DNA libraries were prepared from 3 mg dsDNA using the Agilent SureSelect Human All Exome Kit v.4 (pre-2017) or v.6 (2017 onwards). Sequencing libraries were hybridised with the capture library for 24hrs at 65°C. The hybridised DNA was then captured by streptavidin-coated magnetic beads and amplified with indexing primers. After purification, the amplified DNA was analysed with the Bioanalyzer High Sensitivity DNA Assay. Samples were multiplexed using four samples on each lane, and 100-bp paired end sequencing was performed on the Illumina HiSeq2000 System (pre-2017) or HiSeq3000/4000 (2017 onwards).

2.4.3 Whole genome sequencing

The WGS samples were prepared in batches of 96 and processed using the Illumina TruSeq DNA PCR-Free Sample Preparation kit. The final libraries were checked using the Roche LightCycler 480 II. Samples sequenced with 100bp and 125bp reads utilised three and two lanes of an Illumina HiSeq 2500 instrument, respectively. Samples sequenced with 150bp reads utilised a single lane of a HiSeq X instrument. Following sample and data QC at Illumina, BAM files were received at the University of Cambridge HPC.

2.5 Data processing pipeline for whole exome sequencing

Despite the ability of high-throughput sequencing technologies to cost effectively create large genomic datasets, the multiple steps involved in the sequencing process have the potential to create many sources of artefact and technical variation reducing accuracy of results and limiting data interpretation. To decrease these systematic sequencing errors, a joint variant calling across all samples was performed as well as downstream quality control measures of variants (125).

Within this study, samples were gathered from two related projects that had been sequenced on different platforms, one produced from whole exome and one by whole genome sequencing. This resulted in comparable but differently generated data. Therefore, the two datasets were merged and re-processed using the same in-house variant calling pipeline used by the Simpson group (123) to standardise as many data parameters as possible and minimise the impact of potential artefacts. All 277 WGS raw data files for the SRNS subproject were downloaded from the Cambridge High Performance Computing Service and transferred to the King's College London server in BAM format. Exonic sequences (exons + 10bps) were extracted from the WGS BAM files and reverted back to FASTQ sequences using SAMtools (Sequence Alignment/Map Tools) (113) (Figure 12). FASTQ files were then processed in the same way as the WES samples, using the following steps as shown in Figure 12. Once the dataset was merged, this demonstrated that there were 66 duplicated samples and the data sets adjusted accordingly.

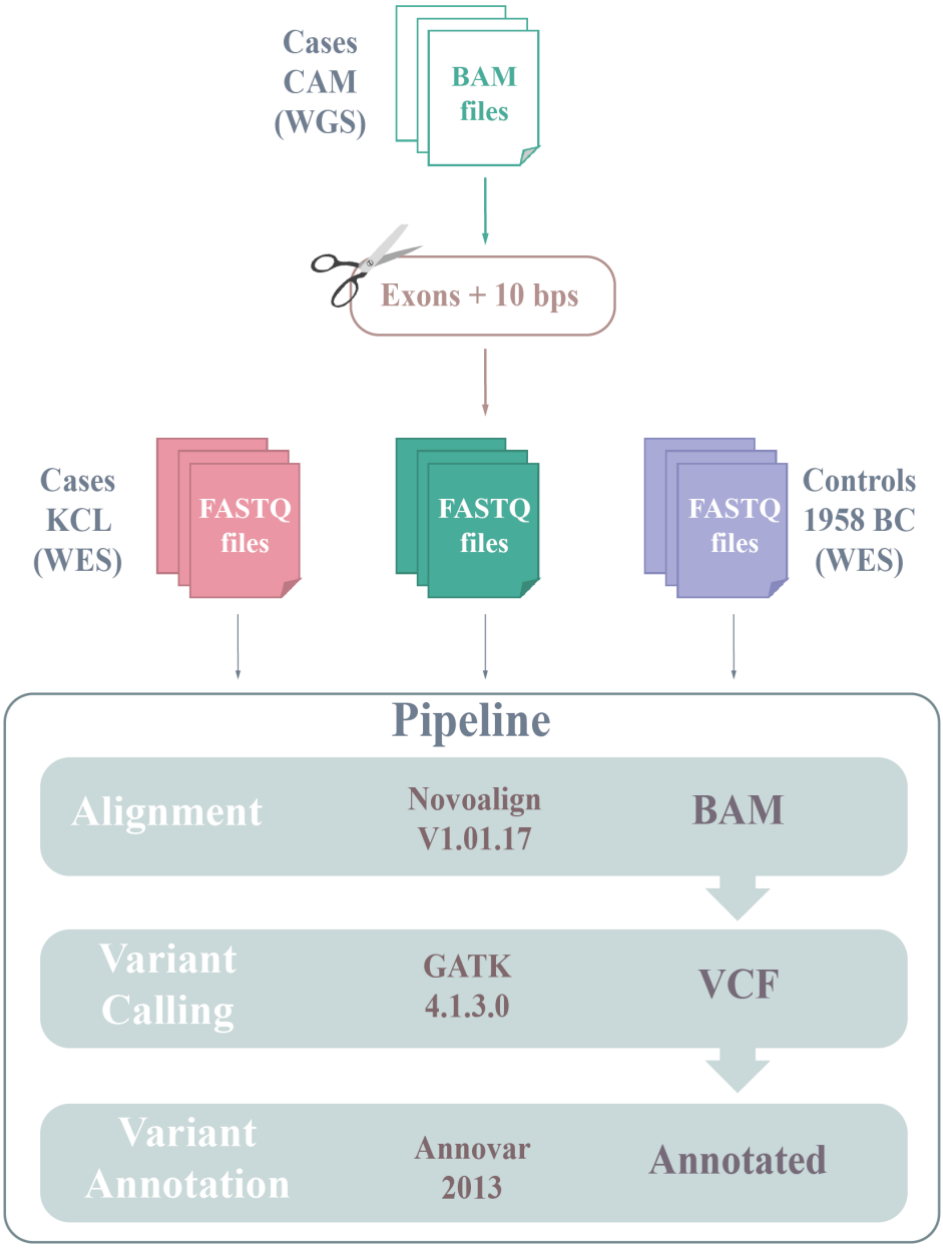


Figure 12. Overview of the data processing pipeline for whole exome sequencing. Three datasets, cases sequenced in King's College London University, cases sequenced in Cambridge University and controls sequenced for the 1958 National Child Development Study, were procressed using the same in-house pipeline. Cases from Cambridge University had to be transform into exome format to be compared with the rest of the samples. Pipeline can be summarised into three main steps: alignment, variant calling and variant annotation.

2.5.1 Alignment

Reads from the FASTQ files were aligned to the reference genome by NovoAlign (http://www.novocraft.com/products/novoalign/) assigning a quality score to each base pair. The human reference genome build used was the GRCh37 version. Two thresholds were applied: gap opening penalty=65 and gap extension penalty=7. After alignment, reads from each sample were stored in BAM file format and sort by coordinate using SortSam tool from the Genome Analysis Tool Kit (GATK 4.1.3.0) (https://gatk.broadinstitute.org/hc/en-us) (126). Duplicate reads arising during sample preparation (PCR artefacts) were removed by MarkDuplicates tool from GATK.

2.5.2 Variant calling

The positions of the reads that differ from reference genome were identified using SAMtools (v0.1.18) (113). The SAMtools algorithm with the '-mpileip' command calls variants in each sample individually and the final data is later combined for the statistical analysis. Several quality metrics were taken into consideration when calculating variants that pass the filtering criteria. Variants were filtered using the 'vcfutils.pl varFilter' command, setting the minimum number of alternate reads supporting each allele to 4. The output data generated after the variant calling was saved in VCF format. The raw VCF files were filtered using VCFtools (v0.1.14) (127). The minimum genotyping quality was set to 20 and minimum read depth to 10.

2.5.3 Joint variant calling

Joint variant calling was also performed across all the samples using the GATK workflow (126). Thus, instead of analysing a BAM file individually, all BAMs (from controls and cases) were analysed in separate batches. Then, all batches were merged in a downstream processing step, gathering genetic information from the whole cohort onto the same VCF file. The file contained information on the total called allele counts

(AN) and alternative allele counts (AC) and quality scores (QUAL) for each of the variants (128). Furthermore, the joint variant calling distinguishes between variants that are not seen in samples because they match the reference genome at the variant location (represent in the genotype column of the VCF file as "0/0") and variants that are not seen because no call is made at that location for technical reasons (represent in the genotype column of the VCF file as "./."). Therefore, by sharing information across all samples, the sensitivity to detect variants increases and the called genotypes are more accurate. This strategy also known as multi-sample calling, aims to compensate for low or missing coverage in some samples and to reduce calling differences that may have arisen due to variants in average sequencing depths.

2.5.3.1 Variant Quality Score Recalibration (VQSR)

All variants from the joint variant calling were then score with the Variant Quality Score Recalibration algorithm generated by GATK (129). VQRS is a sophisticated statistical approach that calculates a new quality score called VQSLOD for each variant by using multiple properties (of the variant context) that are not captured in the QUAL score. This new score gets added to the INFO column in the VCF file. Traditionally, variants are filtered if their values (number of reads covered each allele, proportion of reads forward/reverse, etc) are above or below the set arbitrary thresholds. However, VQSR uses machine learning algorithms (a Gaussian mixture model) that were trained using validated variant resources such as 1000 Genomes and HapMap, to learn from each dataset what is the annotation profile of true genetic variant versus a false positive (usually a sequencing or data processing artefact). Therefore, variants can be filtered to increase sensitivity or specificity depending on the aim of the study (Figure 13). Because of differences in annotation distributions,

VQRS was applied separately for single-nucleotide polymorphisms (SNPs) and indels.

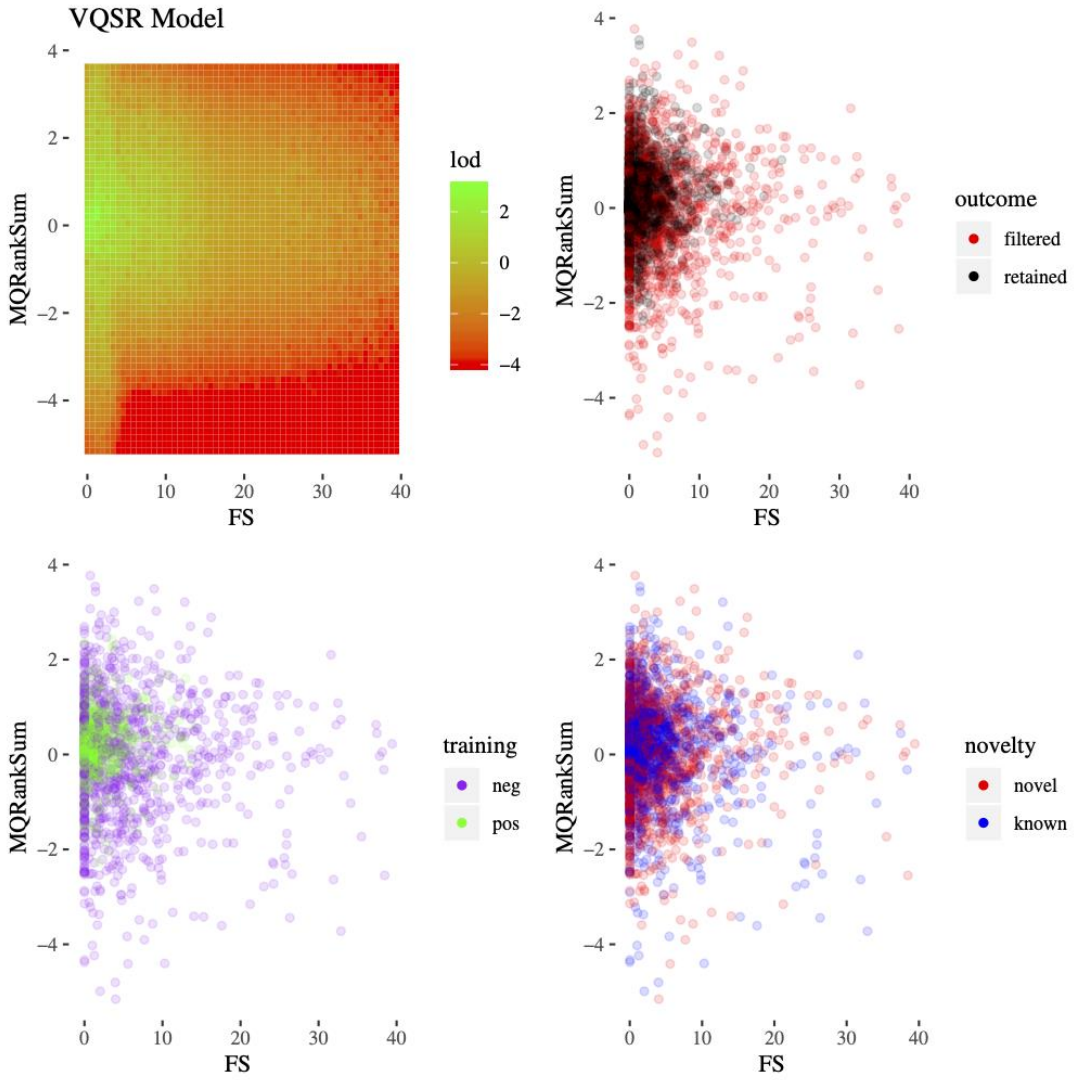


Figure 13. Gaussian mixture model report for SNPs automatically generated by the VQSR tool. Projection of Mapping Quality Rank Sum Test (MQRankSum) versus Fisher Strand (FS). FS is the PHRED-scaled probability that there is strand bias at the site. The upper left plot shows the probability density function that was fit to the data. Green areas are indicative of being high quality (positive LOD) whereas red areas show potential low quality (negative LOD). For the remaining three panels each SNP from the dataset is represented coloured in different ways to show different aspects of the data (outcome, training and novelty). Indels were not included in this report.

2.5.4 Variant annotation

Variants were annotated using ANNOVAR (http://annovar.openbioinformatics.org/) transforming the input file (VCF format) into an annotated variant tab-delimited text file (130). This tool has multiple built-in databases allowing gene annotation, variant type annotations, population frequencies and pathogenicity prediction scores. Variants were annotated with gene, region, variant class, variant type, zygosity and functional consequence including the nucleotide and amino-acid change (Table 3).

Table 3. Variant class annotations

Variants Class	Description
Nonsynonymous	A single nucleotide substitution that leads to an amino acid change
Synonymous	A single nucleotide substitution that leads to the same amino acid being encoded
Stop gain	A single nucleotide substitution that leads to the introduction of a premature stop codon
Stop loss	A single nucleotide substitution that leads to the loss of the wild type stop codon
Splicing	A single nucleotide substitution in the essential splice site, one or two nucleotides adjacent to the splice site
Frameshift indel	An insertion or deletion of several nucleotides that leads to a frameshift of the amino acid sequence
Nonframeshift indel	An insertion or deletion of several nucleotides that leads to the addition or deletion of a number of amino acids

2.5.4.1 Online databases and genome browsers

Variant frequencies for European population were also estimated and annotated from three large datasets of unaffected individuals via ANNOVAR. This included 1,000 genomes from the 1KGP, 6,515 exomes from the NHLBI Exome Sequencing Project (ESP) (131) and 60,706 exomes from the Exome Aggregation Consortium (ExAC)

(132). In downstream QC, allele frequencies from variants of interest were updated using the Genome Aggregation Database (gnomAD) that includes 125,748 exomes and 15,708 genomes (133).

Variants of interest were explored using browsers such as OMIM (Online Mendelian Inheritance in Man) (https://www.omim.org/) and/or ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/) where more than 5,000 monogenic phenotype-gene relationships are recorded.

2.5.4.2 Pathogenicity prediction scores

Pathogenicity prediction tools are routinely used in clinics to prioritise pathogenic variants potentially associated with a specific phenotype over variants of unknown significance. These predictions can be based on genetic, molecular, evolutionary conservation and structural information. These tools estimate whether a genetic variant is damaging, altering the normal levels or function of the protein encoded by a gene, or deleterious, reducing the reproductive fitness of carriers. Regarding sequencing data, the following pathogenicity prediction scores were annotated in the variants via ANNOVAR:

- Polyphen-2 (Polymorphism Phenotyping v2): uses a classifier trained model to predict the consequence of a genetic change. There are two versions of the tool HumDiv and HumVar. Both are based on sequence and protein structure information obtained from UniProt (134). Polyphen-2 score ranges from 0 (benign) to 1 (damaging).
- SIFT (Sorting Intolerant from Tolerant): evaluates the effect of amino acid substitutions (nonsynonymous polymorphisms) in the protein function based on sequence homology and the physical properties of amino acids. It calculates

a score for all possible amino acids at a given position considering that highly conserved residues are more likely to be deleterious. Thus, variants in regions that are highly conserved amongst different species are more likely to have a larger effect on the gene function. SIFT score <0.05 is consider damaging or deleterious (135).

- CADD (Combined Annotation Dependent Deletion): is a framework that integrates a total of 63 annotations from a diverse range of sources into one metric by contrasting variants that survived natural selection. CADD scores are available for 8.6 billion possible human SNVS, including *de novo* mutations. In comparison with other scoring tools, CADD is more informative because instead of focusing on a single information type, it objectively weights and integrates different annotations. CADD score over 20 corresponds to the tip 1%, whereas 30 corresponds to the top 0.1%. The authors of the tool suggest a cut-off of 15 for deleterious variants (136). The version used in this project was CADD v1.6 that includes scores for splicing variants by integrating two deep learning models.

2.6 Data processing pipeline for whole genome sequencing

In chapter 6, common genetic variation predisposing to SRNS is explored studying WGS data that was processed in the pipeline at the University of Cambridge HPC. All protocols used to perform the genotype calling can be found in the recent Nature paper published on whole genome sequencing of rare disease in NHS patients by Ernest Turro et al (102). Samples were aligned by Illumina with the Isaac aligner version SAAC00776.15.01.27 (137) to the human genome build GRCh37. SNVs and small indels were called using the Illumina Starling software version 2.1.4.2. All variants were annotated with deleteriousness scores and conservation scores and with various

summary statistics (allele count, allele number, genotype count, minor allele frequency and call rate). The degree of relatedness between participants and their ancestry were assess using PLINK v1.9 (138). The kinship matrix generated with PLINK was passed to PRIMUS (139) to obtain clusters of related participants.

2.7 Data quality control

Before embarking analysis, quality control steps were performed to assure a high-quality sequencing data and minimise the number of false positive calls both at the individual level and also in the joint-variant calling. Variants located in low complexity regions (LCRs) which are regions of biased composition that are difficult to map because their short repeats of a single amino acid, were excluded from the study following the protocol used by H Li et al (140). The LCRs coordinates used were in BED (Browser Extensible Data) format and can be found at: https://github.com/lh3/varcmp/blob/master/scripts/LCR-hs37d5.bed.gz

For the single individual VCF files, variants were filtered based on the read depth (DP, number of sequencing reads on forward and reverse strand), any genotypes with less than 4 reads were set to missing. Additionally, variants with quality control score (QC) lower than 20 and genotype quality (GQ) lower than 20 were excluded from the analysis.

For the multi-sample calling, VCFtools (v0.1.14) (127) was used for additional quality control of the genotyping data filtering by MAF, call frequency and deviation from Hardy-Weinberg equilibrium (HWE) with different thresholds depending on the analysis (rare or common variant analysis). Global ratio of heterozygous to alternative homozygous alleles (het/hom ratio) was estimated and the outliers were excluded using the R package ‘outliers. An additional QC parameter measured to establish

single nucleotide variant (SNV) accuracy was the transition/transversion ratio (Ts/Tv). Since transitions are changes in nucleotides of similar molecular shape (G↔A or C↔T), whereas transversions are changes in those of different shapes (G↔C, G↔T, A↔T, A↔C), the ratio differs between synonymous and non-synonymous variants, and its distribution varies between different genomic regions and the read length batch. The Ts/Tv ratio measures 2.6 - 3.3 for exome sequence data, with lower ratios suggesting technical artefact (128, 141). The Ts/Tv ratio was consistent across samples (SRNS cases and controls) in the multi-sample calling VCF, and met the expected value for WES data meaning that QC parameters had been met and further analysis could proceed (Figure 14).

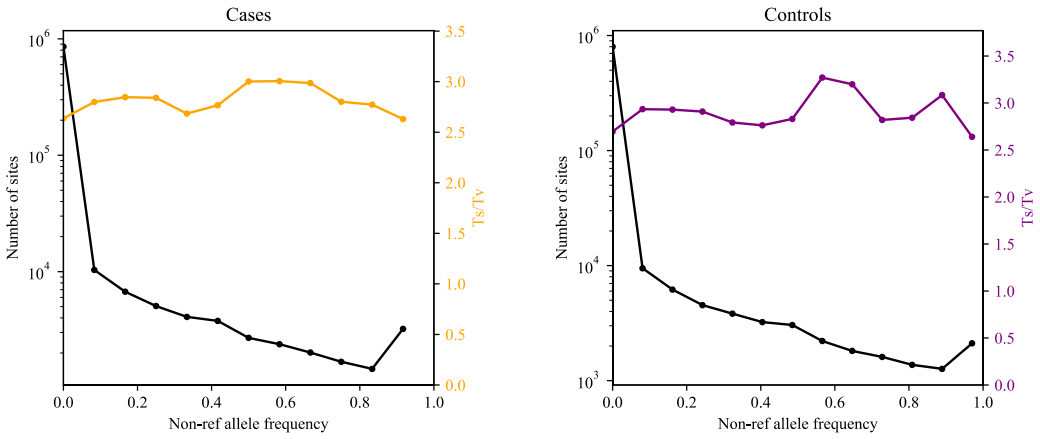


Figure 14. Ts/Tv ratio by AF for cases and controls.Ts/Tv ratio was calculated for all variants using the python package vcftats for cases and controls and shown by the allele frequency.

2.7.1 Relatedness and ancestry

Samples were adjusted by pairwise relatedness using KING (142) to define the maximum number of unrelated individuals and to correct cross-sample contamination due to technical issues during library preparation or sequencing. The pairwise relatedness matrix was generated with KING using the command "--kinship". Kinship coefficient was estimated across all samples including families. A negative kinship

coefficient indicated an unrelated relationship. Close relatives can be identified by the kinship coefficient values: duplicates or monozygotic twins (>0.354), first degree relatives (0,177-0.354), second degree relatives (0.0884-0.177), third degree relatives (0.0442-0.0884) and less than third degree (<0.0442).

All individuals were originally recorded with one of the following self-reported ethnicities: African, East Asian, European, Other and South Asian. However, to ensure this was accurate, population substructure identification was performed on all cases by KING with the multidimensional scaling option (MDS) to confirm reported ethnicity. MDS calculates by default 20 principal components (ancestry coordinates) for each sample using a subset of common independent exonic variants. Samples were projected onto the first (PC1) and second (PC2) principal components and compared with a reference population from 1KGP to identify their ethnicity.

2.8 HLA typing from WES and WGS

Since there is extensive linkage disequilibrium and great allelic differences present in the MHC locus, most of data processing pipelines are unable to generate accurate genotype calls within the region. As a consequence, the majority of the short sequencing reads from HLA genes are either not mapped correctly or fail the alignment process generating a set of unmapped reads subsequently excluded from analysis. Accordingly, multiple methods have been developed to correctly classify the sequencing reads from HLA genes by using different alignment strategies and comprehensive reference panels. Specifically, there are tools that perform HLA typing using high-throughput sequencing data which count on the unmapped reads from the samples and validated reference datasets such as the International HapMap Project

(143) 1KGP to improve the sequencing of the region. The approach used in this study for both WES and WGS is summarised in Figure 15.

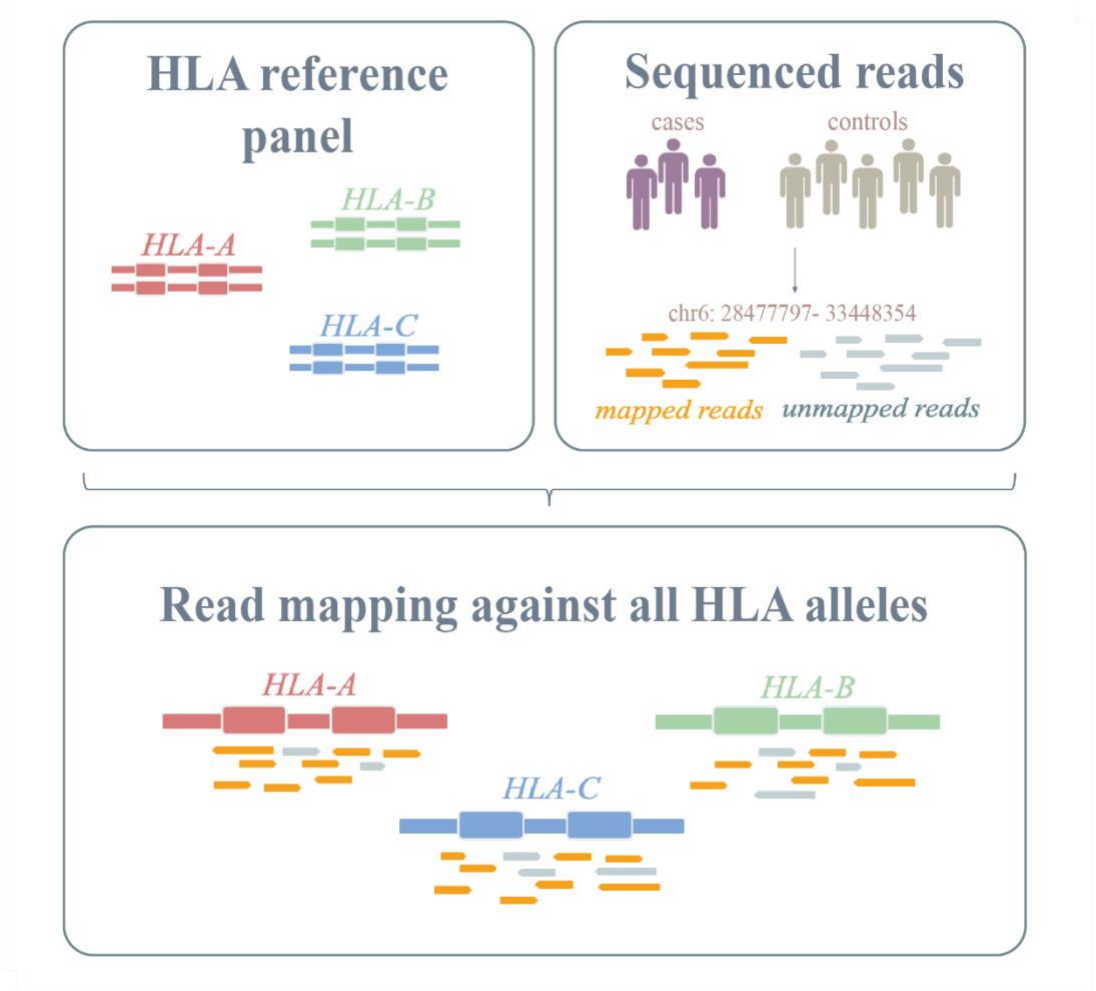


Figure 15. Overview of data analysis for HLA typing. Sequenced reads (mapped and unmapped) from WES and WGS samples are aligned to a comprehensive HLA reference panel to search for best matching alleles based on alignment statistict.

Furthermore, through their exceptional polymorphism, alleles of the HLA genes have been represented at the protein level (using serological techniques) with a specific nomenclature introduced in the 1987 Nomenclature Report (144). HLA alleles are defined by the name of the gene follow by sets of digits separated by colons (i.e.: HLA-DQB1*06:04). The first two digits describe the allele family that encode the serologically defined antigen. The third and fourth digits are the specific HLA protein. Since the first proposal in 1987 more digits have been added up to eight digits

depending on the sequence (145). Here, HLA typing tools used HLA nomenclature and alleles were estimated with four digits resolution.

2.8.1 Benchmarking of methods for HLA typing

For HLA typing, samples were aligned using NovoAlign and Burrow-Wheeler Aligner (BWA) (146). All reads from the HLA region (chr6:28,477,797-33,448,354) were extracted from the BAM files by SAMtools. Unmapped reads were selected using SAMtools with the command 'view -f 4' for unmapped segments. HLA-VBSeq and HLA-Genotyper were chosen to estimate HLA alleles because both were specifically built to perform HLA typing from WGS and WES data as input. Estimation of HLA types by HLA-VBSeq was performed with the HLA v2 database based on IMGT/HLA database Release 3.31.0. HLA-Genotyper was used to call HLA genotypes with 4-digit resolution with the options '--genome' and '--exome' accordingly. Additionally, to improve the speed and accuracy of the HLA genotype calls, ethnicity of the samples (EUR) was specified. This option provides a set of priors for the various HLA alleles found in the ethnic population supported by HLA frequencies found in 1KGP.

A typing accuracy comparison between HLA-VBSeq and HLA-Genotyper was also performed to ensure accuracy. Both tools estimated the HLA types of 22 European samples that underwent WES and WGS. The genotype called for two allele groups, HLA-DQA1 and HLA-DQB1 were compared across sequencing platforms and methods. Since HLA-Genotyper gave better results than HLA-VBSeq when concordance of alleles was compared from the same samples sequenced in different platforms, this was subsequently used for experiments (Figure 16).

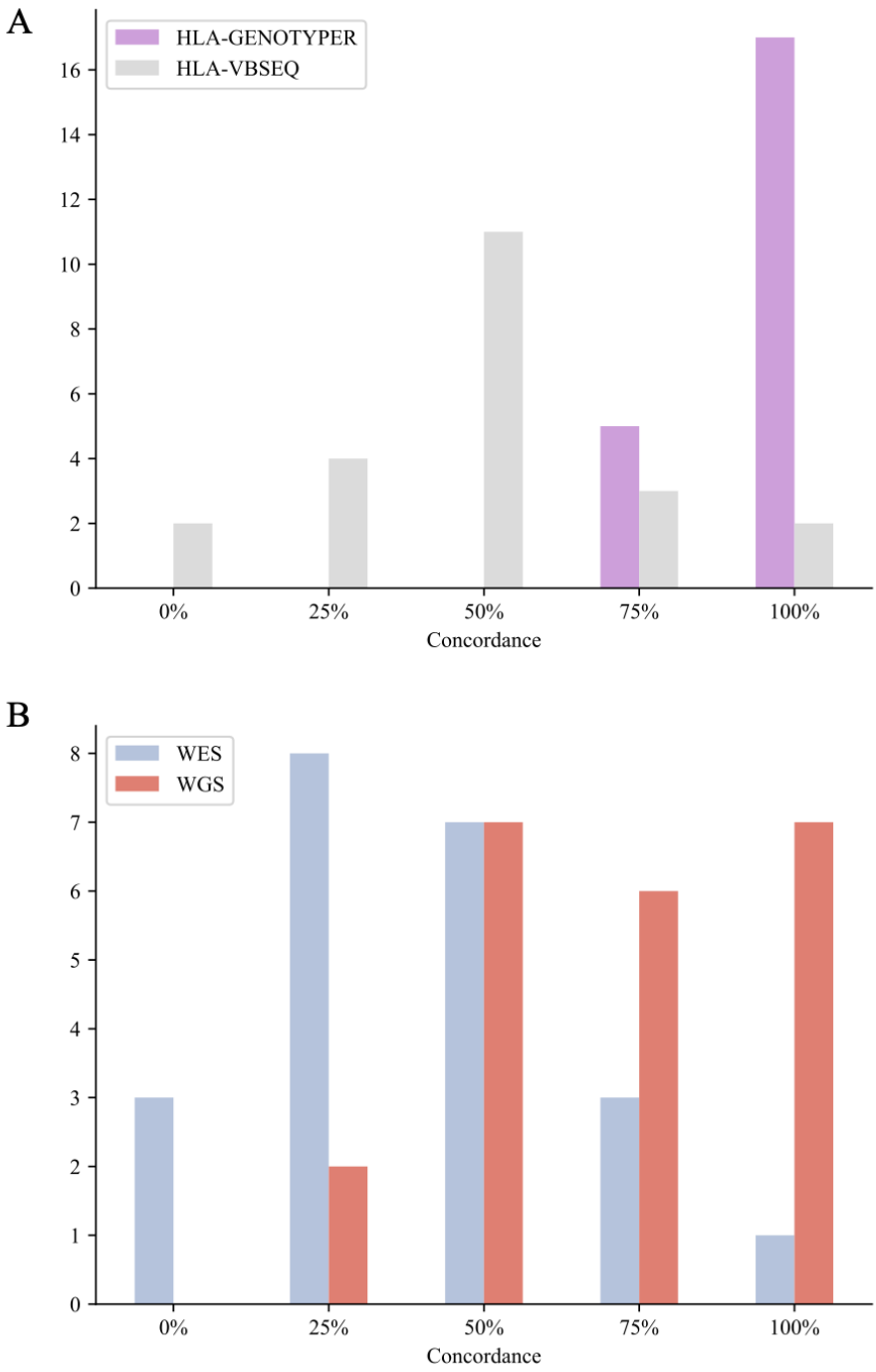


Figure 16. HLA typing accuracy comparison. HLA alleles for DQA1 and DQB1 of 22 European samples that underwent whole exome and whole-genome sequencing were compared. (A) Comparison of two different HLA typing methodologies (HLA-Genotyper and HLA-VBSeq) across different sequencing technologies. HLA alleles estimated in WES and WGS samples were compared by two different typing tools HLA-Genotyper (purple) and HLA-VBSeq (grey) with overall concordance 94.3% and 48.8% respectively. (B) Comparison of the HLA typing methodologies using the same sequencing platform. HLA alleles estimated by HLA-Genotyper and HLA-VBSeq using WES data were compared (blue) with overall concordance of 39.7%. HLA alleles estimated by HLA-Genotyper and HLA-VBSeq using WGS data were compared (red) with overall concordance of 70.45%.

2.9 Computational and statistical approaches to identify disease-causing variants using whole exome and whole genome sequencing.

Multiple study designs and analytical strategies were carried out using high-throughput sequencing technology to identify causal genetic variation and to better understand the genetic basis of rare diseases (Figure 17). A clustering of disease within families suggests genetic and/or share environmental risk factors. Primary SRNS and in some instances SSNS can be considered Mendelian diseases in view of clear monogenic inheritance patterns within families. Since a number of such families were recruited, pedigree information was used to narrow down the search for candidate causal alleles in family studies. Additionally, SRNS and SSNS also present characteristics of complex disease, with considerable heterogeneity in terms of the genetic architecture with potentially polygenic forms. Thus, strategies to detect common disease-causing variants were also used.

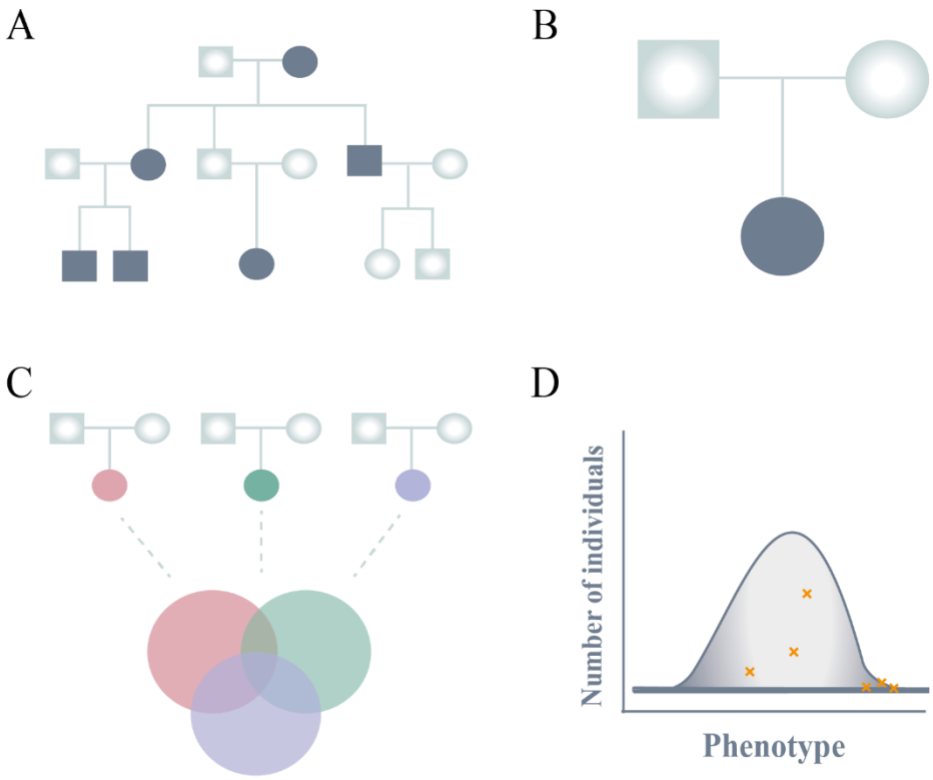


Figure 17. Strategies for finding disease-causing variants using high-throughput sequencing. (A) Linkage analysis to search for a shared region of the genome segregating within the affected individuals and not present in the unaffected. (B) Sequencing of unaffected parents and affected child (trios) for identification of de novo conding mutations. (C) Filtering variants across unrelated affected individuals to indentify variants within the same gene or genes. (D) Individuals with rare variants in the same gene sharing an extreme phenotype of the disease. Figure adapted from MJ Bamshad et al (2011).

2.9.1 Family-based study designs

Family analyses evaluate genetic markers searching across the entire genome for regions harbouring potentially causal risk factors. Historically, family-based studies have been the first approach to detect genes responsible in monogenic disorders (147). Segregation and linkage studies have been particularly successful in cloning highly penetrant diseases causing genes. By studying relatives in the same family with the phenotype of interest, the genomic search space is substantially restricted increasing the power for gene discovery. This allows identification of rare and probably damaging variants shared by the affected members and not present in the unaffected members. The number of relatives studied can range from two family members to big

pedigrees with multiple generations. Therefore, sequencing of unaffected parents and affected child (trios), were considered sporadic cases and studied for identification of *de novo* coding mutations instead of being analysed with any of the family-based strategies described in this section.

2.9.1.1 Segregation analysis

Segregation analysis uses statistical methodology that attempts to determine if the phenotypic pattern within a family is consistent with the genetic inheritance of a disease, specifically with a view of identifying single gene defects. This technique can be performed using family sequencing data where large pedigrees with many affected members are particularly informative to identify genes linked to a disease. Thus, selecting enough affected individuals to yield sufficient numbers is crucial. Moreover, in this type of analysis some assumptions are made about the underlying mode of inheritance (i.e. dominant or recessive inheritance) and the impact of environmental factors on a given trait. Depending on the family history and mode of inheritance, different parameters were changed accordingly to find the model of best fit for the family data. Equally, if there was insufficient information about the family history or the pedigree was too small to securely identify a model of inheritance, the stringency of filters used for searching for segregation in the affected individuals were adjusted accordingly.

All families in this project were studied through segregation analysis by extracting a list of variants shared among the affected individuals that were not present in the unaffected. Families without a clear pattern of inheritance were studied using different models. Variants selected for dominant model of inheritance were heterozygous whereas for recessive model of inheritance were homozygous or compound

heterozygous when two or more heterozygous alleles at a particular gene locus (one on each chromosome of a pair) were found. The list of variants segregating within the affected individuals of each family was filtered by allele frequency using gnomAD (133) annotation (section 2.5.4.1) (AF<0.01 for recessive model and AF<0.001 for dominant model). For families that underwent WES, the segregation analysis was restricted to coding variants. For families that underwent WGS, the segregation analysis was performed across all variants including non-coding ones and prioritise by consequence on the protein level. Additionally, for the WGS samples variants overlapping with low-complexity regions were excluded from the analysis.

2.9.1.2 Parametric linkage analysis

Parametric linkage analysis is a method that identifies regions of the genome underlying a given trait by testing a set of markers (alleles) for cosegregation with disease status within either a family or across several families. Markers located close together on the same chromosome are more likely to be coinherited than would be expected by chance, because their proximity means that recombination is less likely to separate them. Linkage mapping use a set of markers evenly distributed across the genome (every 10 cM), to capture most of the recombination events. Linkage studies are considered parametric when the data is analysed assuming a specific genetic model. This strategy has been successful in detecting rare and highly penetrant disease-causing variants with unusually large effect sizes that do not impact on reproductive fitness.

The identification of a locus that co-segregates with disease status was confirmed statistically by the LOD score, defined as the ‘logarithm of the odds’. This determines the likelihood of whether co-segregation occurred by chance or that the disease locus

and markers are co-segregating assuming a particular recombination. A LOD score calculation is therefore used to measure the likelihood that markers and disease are inheritance non-independently. In 1955, Morton derived the LOD score criterion of 3, with the goal of obtaining a posterior false-positive rate among reported linkages that was <5%. The value of 3 reflects the fact that the prior probability that two randomly selected loci are syntenic and within reasonable mapping distance (say, recombination fraction [0] < .3) is small, on the order of 2% (148). In practice, a LOD score below 3 it is provisional and further evidence is required to confirm linkage. A LOD score of less than -2 is considered good evidence that two loci are independent and are therefore used for exclusion of linkage.

One family, Family A, had sufficient members spanning 3 generations to perform parametric linkage analysis. The observed inheritance of the trait for Family A was consistent with autosomal dominant inheritance. 12 family members underwent WES (7 affected and 5 unaffected) and 4 of those members also underwent WGS. A multi-sample calling with WES data was performed across all 12 members of the family to create a joint VCF file. All genotypes with a quality below 15 were excluded. The minimum number of sequencing reads on forward and reverse strand for any variant was set to 10 and the maximum to 1000. All had to be genotyped in at least 90% of total individuals. The VCF file was converted to linkage format (.PED, .MAP and .DAT) by PLINK. The parametric linkage analysis was performed using MERLIN (Multipoint Engine for Rapid Likelihood Inference) (149) software, under the assumption of autosomal dominant inheritance. MERLIN's linkage parametric analysis is based on the Lander-Green algorithm (150) and creates sparse inheritance trees for pedigree analysis. The measures from the Lander-Green algorithm depend on the pedigree size and the number of markers, being linear for the number of loci and

exponential in the number of individuals in the family. The first step was to verify the input files (pedigree and data file) using the option pedstats. Then, error detection was performed with the option '--error' to identify genotyping error that can lead to misleading inferences. The genotypes that were flagged during error detection were reject from the analysis using pedwipe. Finally, a parametric linkage analysis was carried out generating rapid haplotyping with genotype error detection and LOD scores for each marker. For this analysis, the file 'parametric.model' specified the parameters (Table 4) that best fit the family data. The disease allele frequency was based on the prevalence of SRNS and the penetrances reflect an autosomal dominant model with a low phenocopy rate for non-carriers. Parameters specified in MERLIN were as followed:

Table 4. Parametric model used in MERLIN.

Affection	Disease Allele Frequency	Penetrances	Model Name
SRNS	0.0001	0.0001,1.0,1.0	Rare_Dominant

2.9.1.3 Nonparametric linkage analysis

Nonparametric linkage analysis also known as model-free does not require any genetic model of inheritance to be specified and is normally applied for the study of complex diseases. This approach relies on the assumption that affected individuals in the same family will share markers or chromosomal regions, regardless of the mode of inheritance. This methodology uses marker data in affected siblings from multiple families in order to identify markers that have been shared among families more often that would be predicted by random Mendelian segregation. Siblings are expected to share 50% of their genes. However, if there is cosegregation, affected siblings will share more alleles identical-by descent (IBD) in the region of interest than as might be expected by chance. The probability that a specific marker allele is cosegregating with

disease status, with recombination fraction θ, is compared (likelihood of odds) with the probability that the marker and disease status are not cosegregating (θ=0.5). Nonparametric linkage analyses require higher thresholds for statistical significance. LOD scores above 5.4 are considered highly significant evidence for linkage, between 3.6 and 5.4 are considered significant and above 2.2 are suggestive (151).

A nonparametric linkage analysis was performed in 7 informative families where no causal variant was found in the coding regions of established SRNS genes (Family 1, family 2, family 4, family A, family D, family E and family F). For a family to be informative, at least two affected children within a generation are necessary. A multi-sample calling with WES data was performed across all families to selecting common and good quality genetic markers. All genotypes with a quality below 30 were excluded. The minimum number of sequencing reads on forward and reverse strands for any variant was set to 20. Additionally, common variants were selected filtering by allele frequency (AF>0.2). The VCF file was converted to linkage format (.PED, .MAP and .DAT) by PLINK. The first step verified input files (pedigree and data file) using the option '--pedstats'. Nonparametric linkage analysis was performed by MERLIN with both options the Whittemore and Halpern NPL '--pairs' and NPL all '--npl' statistics. The standard nonparametric linkage analysis performed by MERLN is based on the Kong and Cox (152) linear model that evaluates the evidence for linkage. However, for this analysis the '--exp' option was used to search for a large increase in allele sharing in a small number of families providing a more accurate and sensitive test for linkage.

2.9.2 Case-control association studies

Case-control association studies use genotype data from individuals with the same phenotype to identify alleles that are found in greater or lower frequency compared to a group of unaffected controls. These studies require the identification of unrelated individuals with the same phenotype, which can be particularly challenging for rare diseases through a lack of sufficient numbers of cases, and the identification of controls without the phenotype. Furthermore, appropriate numbers of cases and controls are crucial to reach enough statistical power.

Multiple strategies are available for case-control analysis to enable testing for genetic association such as single variant tests, multiple variant tests, collapsing methods and aggregation methods (114). Out of these, two were chosen to test for genetic variation within the cohort: genes-based burden test and genome-wide association study. The gene-based burden test evaluates the implication of rare genetic variation at gene-level and all variants across a unit (in this case a gene) are collapsed together to increase power. Although rare variants individually are infrequent, together might be presented in sufficient frequency to be compared against unaffected controls. In contrast, GWAS evaluates common variants by testing a single variant at a time. This single variant test is more powerful for common variants than for rare variants considering identical effect sizes through the number of rare variants being higher than the number of common variants (153).

2.9.2.1 Gene-based burden test

Burden testing is a collapsing method where information from multiple rare genetic variants is aggregated into a single count (the total number of variants) and tested for association with a trait. Given the nature of rare disorders, a causal gene might have several unique variants (only found in one individual) enriched in cases compared to

controls, and the statistical differences observed are then more likely to have biological significance once corrected for false positives. Thus, instead of examining whether there is an enrichment of any one variant in cases versus controls, it analyses whether there was an enrichment of variants in any one gene in cases versus controls. For this case-control study, SRNS patients were compared to the 1958 British birth cohort as population control group (124). Analysis was carried out using bash scripts and a software package called EPACTS (Efficient and Parallelizable Association Container Toolbox) (https://genome.sph.umich.edu/wiki/EPACTS).

Any systematic differences between sequencing data from different platforms is likely to introduce bias in association testing. Therefore, to minimize technical artefacts, the case-control analysis was performed using the multi-sample calling. By performing variant calling jointly, if a variant is called in a set of samples, the multi-sample calling checks whether a call (reference or alternative) is made at this position for all samples. Analysis was restricted to set of variants that have been genotyped in at least 90% of individuals. Any sample with a high number of 'missingness' (>0.25) where a genotype could not be called was removed. Additionally, regions that are known to generate false positives because of highly variable genes. The highly polymorphic regions can be found in the work made by KVF Fajardo et al (154). Only cases and controls that clustered via principal component analysis with individuals of European ancestry were selected for study. Samples were also adjusted by pairwise relatedness using KING to define the maximum number of unrelated individuals.

For the burden testing, all variants observed in a gene were aggregated and classified into three categories: alteration (non-synonymous codon-gain and codon loss), truncation (frameshift, introduction of stop-codon and alteration of splice site) and

synonymous. Variants were further subdivided into two groups (dominant and recessive) and filtered by minor allele frequency; for a dominant model, the MAF was set at <0.001 whereas for recessive, MAF was set at <0.01. The frequency threshold used also took into account the disease prevalence. Two other groups of variants were created with a CADD score of more than 10, for both dominant and recessive model. Specific variants and genes (such as known false positives) were excluded by remaking the VCF file eliminating those SNPs or genes and rerunning the EPACTS analysis. Evaluation of the burden association was determined with a one-tailed Fisher exact test.

2.9.2.3 Genome-wide association study (GWAS)

A genome-wide association study was performed to identify whether there was any association between common variants (MAF>0.05) and SRNS. Each variant was tested for association with the phenotype using logistic regression (155), which is able to adjust for potential confounding variables such as ethnicity by using the first four principal components as covariates. The analysis was carried out with EPACTS using the logistic Wald association option. The test used binary phenotypes and collapsed variables (genotypes) with joint estimation covariates implemented by Hyun Min Kang. Thus, phenotypes were converted to two different numeric values (binary) where "0" represented unaffected status and "1" represented affected status for the trait. Wald test is one of the classical approaches to hypothesis testing that only requires the estimation of the unrestricted model, and has the advantage of lowering the computation burden compared with other strategies. In this analysis the null hypothesis was that there is no association between the odds of having SRNS and a genetic variant in the population of interest. The logistic regression model used was as followed, where Y is the expected value of the phenotype, given genotype X and

covariates A,B,C and D (first four principal components). P-values were calculated based on whether β_1 significantly differs from zero.

$$\log\left(\frac{p(Y)}{1-p(Y)}\right) = \beta_0 + \beta_1 X + \beta_2 A + \beta_3 B + \beta_4 C + \beta_5 D$$

In the main, GWAS studies rely on imputing data from genotyping arrays to obtain a prediction of the whole genome sequence of all samples. However here, the imputation step was not required since the SRNS cohort had been whole genome sequenced as part of the Rare Disease Pilot study associated with the 100,000 Genomes Project (102), and the raw genomic sequence could be sourced and used for analysis. Controls also whole genome sequenced using the same platform and pipeline were carefully selected from a comparable population taking into consideration ethnicity which was restricted to European (section 2.6). Ancestry outliers were detected by principal component analysis and excluded from downstream analysis. To minimize technical artefacts, a joint variant calling was performed, with variants excluded from analysis if the call rate was <0.90, or a minor allele frequency was <0.05, or there was deviation from Hardy-Weinberg equilibrium (P <1 x 10-6). Additionally, a genotype quality (GQ) threshold of 30 and depth (DP) threshold of 20 were set per genotype. All sites with a filter flag other than 'PASS' in the joint variants calling were also removed. Genetic relatedness was assessed using a subset of high-quality independent common variants and a maximum unrelated set of samples was generated (PI HAT< 0.09375). Second degree relationships or closer were removed. Following quality control, a total of 3,944,568 genotyped variants were retained and used in the association analyses.

2.10 Other statistical methods

2.10.1 Fisher's exact test

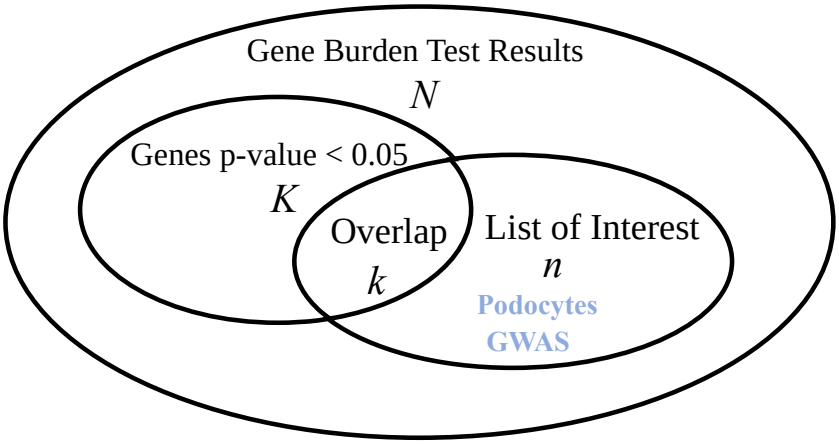
Fisher's exact test was used to assess the results from the gene-based burden test (case control study) described in section 2.9.2.1. It evaluates the independence between two nominal variables, disease status and number of variants within a gene, and studies whether the proportions of one variable are different depending on the value of the other variable (Table 5). One tailed test was selected due to the expectation of an enrichment rather than deficit in the number of variants in cases compared to controls. Therefore, the null hypothesis was that the proportion of rare variants in cases was the same as controls. Fisher's exact test is appropriate for rare variants association studies because the expected sample size is small.

Table 5. Fisher Exact Test contingency table.

	Alleles		
Disease Status	Alt	Ref	Total
Case	x	$m - x$	m
Control	k-x	$n - (k - x)$	n
Total	k	$(m + n - k)$	$m + n$

2.10.2 Hypergeometric test

Hypergeometric testing was used to model the association between two independent gene sets and to calculate the probability of a certain number of genes overlapping between them. This type of test is also equivalent to one-tailed Fisher's exact test. The hypergeometric model used was as shown in the Figure 18, where k is the overlap between the two datasets, K in the genes with a p-value lower than 0.05 in the burden test, n is the number of genes from a list of interest (in this case podocyte enriched genes and CKD associated genes) and N the total number of genes tested in the burden test. The test was implemented by R with the built-in function called 'phyper'.



$$P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

Figure 18.Hypergeometric test diagram and equation.

2.10.3 Dosage analysis

A dosage-based test was performed to confirm an association between SRNS and HLA alleles. Through typing, HLA alleles were estimated for the cases and controls from the GWAS (and replication cohort) and transformed into a dosage file. Number of copies for each HLA gene were represented as followed: 0, 1 and 2. The test was performed with PLINK v1.9 using the '--logistic' and '--dosage' options. As mentioned previously, the first four principal components of cases and controls were included as covariates in the model.

2.10.4 Linear regression

The relationship between SRNS subphenotypes (gender, SRNS type, histology and age-of-onset) and *HLA-DQA1*01:02* genotype was examined using linear regression analysis. Gender, SRNS type (primary and secondary) and histology (FSGS and MCD) information was available for the GWAS cases, and it was transformed into categorical variables. Age-of-onset was available in years as a numeric variable. *HLA-*

*DQA1*01:02* genotype was transformed to a binary outcome: 0 (no copy of the allele) and 1 (one or two copies of the allele). Linear analysis was performed with the R package stats (3.6.0) using the 'lm()' function. The first four principal components of cases were included as covariates in the model.

2.10.5 Colocalisation

Colocalisation evaluates whether or not the same variant (or variants) in a given region are responsible for two potentially related phenotypes. This methodology used the summary statistics from GWAS and expression quantitative trail loci (eQTL) data from a specific tissue of interest.

Estimation of the colocalisation between SRNS signals and kidney cis-eQTLs from GTEx was performed excluding the HLA region and focusing on the suggestive peaks from the GWAS. Kidney eQTL data was downloaded from release V8 that was based on WGS from 73 (kidney cortex) donors, which all had RNA-seq data available (https://console.cloud.google.com/storage/browser/gtex-resources). Candidate kidney cortex eQTLs from European samples were selected from any variant located within the SRNS suggestive risk locus ($P<1x10^{-5}$) or nearby (± 1 Mb from the region of interest). The coloc R package (156) was used to performed genetic colocalization between SRNS association signals and the kidney cortex eQTL signals using a set of variants that overlapped between both datasets. The package is based on a Bayesian model that calculates posterior probabilities of different causal variant configurations under the assumption of a single causal variant for each trait (156).

2.10.6 Multiple testing corrections

Bonferroni correction was used when required in several different analytical processes during this study. This is a multiple-comparison correction used when multiple

dependent or independent statistical tests are conducted simultaneously. The alpha level, which is the probability of rejecting the null hypothesis (also known as significance), might be appropriate for each individual test but not for a set of tests. Thus, Bonferroni correction is the division of the alpha by the number of tests performed.

For the results obtained by the gene-based burden test, the significance threshold of $p<2.5x10^{-6}$ was used. The threshold was calculated by Bonferroni correction of the effective number of independent test (approximately 20,000 genes) assuming $\alpha=0.05$.

For the results obtained by GWAS, the standard genome wide significance threshold of $p < 5x10^{-8}$ was used. Given the linkage disequilibrium and structure of the genome it has been estimated that there are approximately one million independent loci. Bonferroni correction for the effective number of independent test (one million markers) assuming $\alpha=0.05$ has led to the standard genome wide significance threshold limiting false positive associations (157).

Chapter 3 – Phenotypic description of the patient cohort

3.1 Introduction

High throughput sequencing technologies in the form of WES and WGS have become a routine component of gene discovery studies. With this however, come exhaustive categories of genetic variation, some causal and therefore relevant and others of uncertain significance. This emphasises the importance of accurate phenotype-genotype correlation, particularly when examining rare disease cohorts that are generally small in comparison with common disease cohorts where some untoward variation becomes less significant simply through the large number of cases under test.

Consequently, optimising phenotypes by precise annotation raises the accuracy of findings, as it potentially increases power to localise genes of interest and also aids the interpretation of associations between variants and disease outcomes. Phenotype optimisation includes the use of parameters such as symptoms or ages of onset to reduce genetic heterogeneity within a set of cases, allowing analysis of related phenotypes, as well as derivation of new phenotypes. New opportunities are also presented by technological advances that permit efficient collection of phenotypes on an individual conferring maximum advantage for genotype-phonotype correlation and accurate association studies through more detailed phenome data (158).

This chapter explores characterisation of the SRNS study cohort (children and adults, recruited nationally from UK nephrology centres), as well as data collection to allow rigorous definition of the phenotype in each of the cases recruited.

3.2 Cohort description

The main inclusion criteria was diagnosis of primary nephrotic syndrome in children and adults principally primary and secondary SRNS. Probands were identified and recruited together with affected and/or unaffected relatives if possible to form at minimum a duo or trio via local recruitment and two national initiatives: the BRIDGE consortium (https://ega-archive.org/studies/EGAS00001001012) and Radar/NephroS (https://renal.org/rare-renal/patient/nephrotic-syndrome-0). Patients were drawn in the main from Evelina London, Guys and St Thomas' and Bristol as decribed, but recruitment was also from other national UK Paediatric and Adult Nephrology units.

Venous blood was drawn on each case, DNA extracted and then stored by each clinical genetics laboratory hub prior to sequencing by whole exome or whole genome. Additional to genomics analyses, cases were deeply phenotyped. Each had a detailed medical, family and medication history taken as well as inclusion of renal biopsy findings and results of any other relevant clinical investigations. Clinical data was extracted directly from patient records together with sequential data from the UK Rare Renal Disease Registry (http://rarerenal.org/radar-registry) to monitor clinical outcomes. All patients had long term follow up spanning between 5 and 50 years.

The cohort is comprised of two independent datasets; 267 samples that were whole-exome sequenced via GSTT/KCL BRC and 277 whole-genome sequenced as part of the BRIDGE consortium at the Cambridge BioResource. Both sequencing projects were performed on Illumina platforms. However, in view of the differences between WES and WGS as well as degree of coverage, sequencing data on both cohorts was subsequently merged prior to further analysis in order to reduce noise and ensure the data was comparable. Both datasets were merged and processed using the in-house

variant calling pipeline explained in section 2.5. The study cohort consisted of 544 individuals in total, with 66 duplicates (the overlap between the two original datasets), 24 affected family members and 32 unaffected relatives. Thus, the total number of sporadic cases and probands was 422 excluding duplicates and affected/unaffected family members.

3.2.1 Sex ratio

Multiple studies have reported nephrotic syndrome to be more common in males than females, with a male-female ratio in SRNS cohorts of 1.3:1 (82) or 1.4:1 (159). The cohort studied in this dissertation comprises 422 unique cases (excluding duplicates and family relatives) that were originally annotated with a self-reported sex (male or females). Of the 422 patients, 230 (55%) were male and 192 (45%) were female. If affected family relatives are included the cohort includes 446 individuals, 241 (54%) were male and 205 (46%) were female. The male to female ratio of this cohort was 1.19:1, which is broadly consistent with the epidemiology reported for SRNS, but the sex difference was not as strong as in other studies.

Genetic sex was also evaluated based on the number of heterozygous variants in chromosome X in the raw VCF files. A histogram plot of the number of heterozygous variants for all samples confirmed the bimodal distribution with individuals with less than 200 variants being male and with more female (Figure 19).

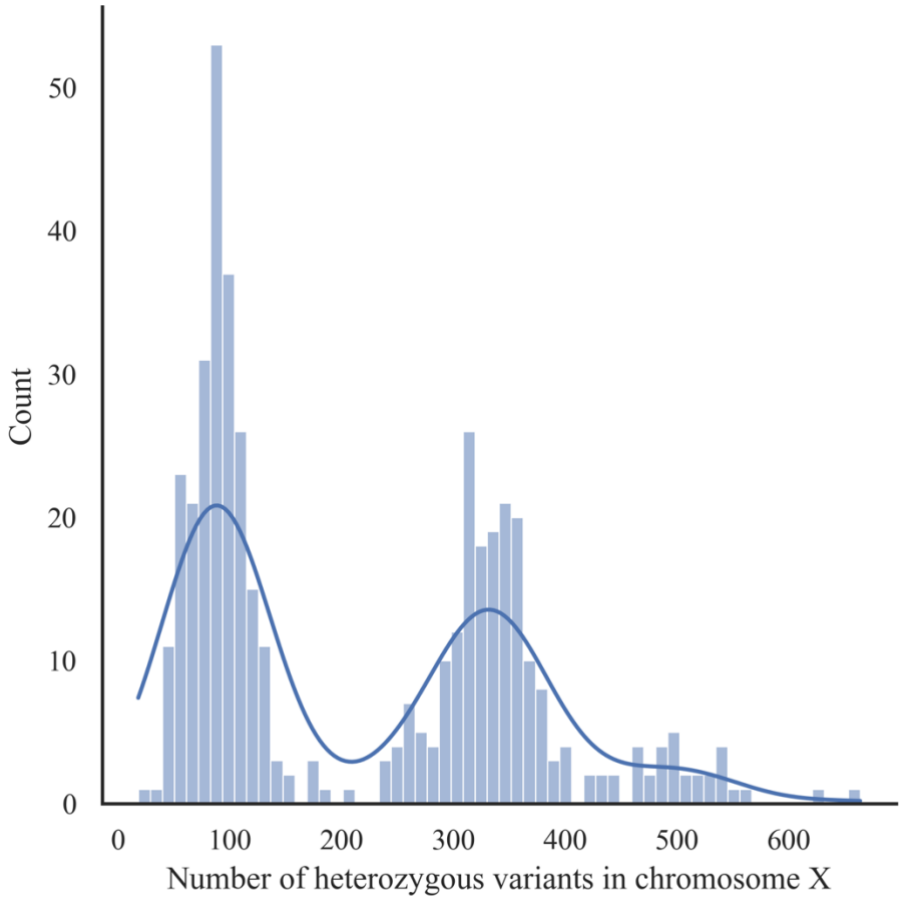


Figure 19. Distribution of the number of heterozygous variants in chromosome X per sample. The number of variants in chromosome X provides information about the sex of the sample. Number of heterozygous variants in chromosome X ranged from 0 to 200 for males whereas for females ranged from 200 to 600.

3.2.2 Ancestry

All individuals were originally recorded with one of the following self-reported ethnicities: African, East Asian, European, South Asian and mixed ancestry. The ethnicity of 478 samples (removing duplicates and including family relatives) was also evaluated using principal component analysis (PCA) on a subset of common (MAF >0.05) independent biallelic exonic variants that did not deviate significantly from the Hardy-Weinberg equilibrium (P >1 x 10-6). Samples were projected onto the first (PC1) and second (PC2) principal components and compared with a reference population from 1000 Genomes Project Phase 3 (103). PCA analysis reveals 68.72% of the study participants cluster closely with individuals of European ancestry, 13.47% with individuals of South Asian ancestry, 7.76% African, 1.59% East Asian and 8.44% do not cluster closely with any of the populations evaluated and likely represented individuals of mixed ancestry (Figure 20).

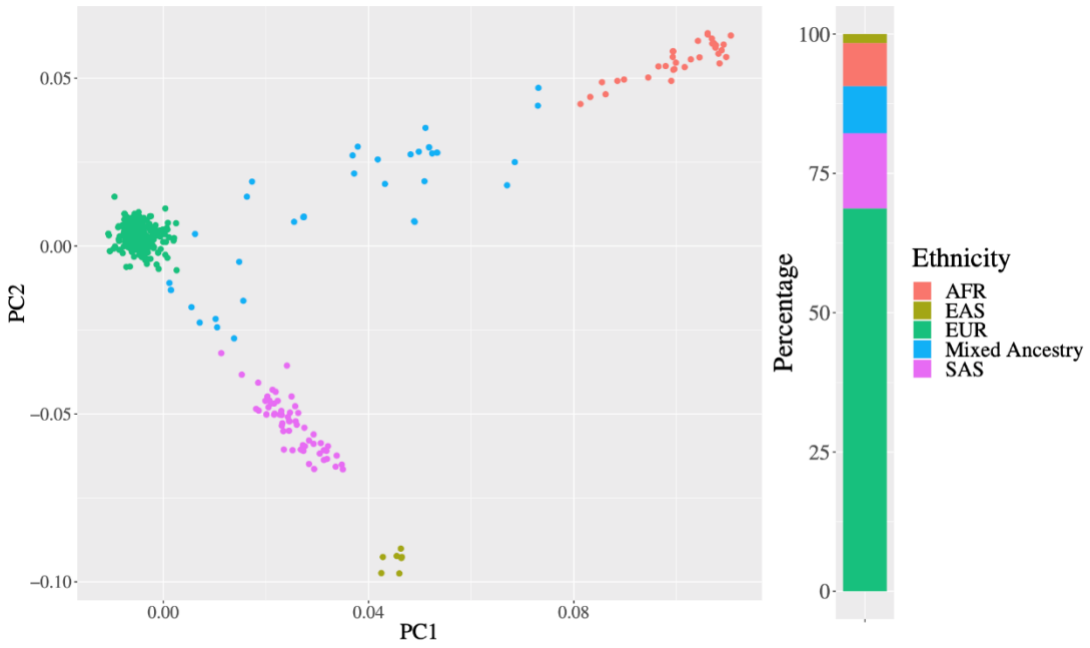


Figure 20. Principal component analysis of the SRNS cohort and percentagesOn the left, projection of individuals onto the first two principal components of genetic variation. Africans are represented in red, East Asians in yellow, Europeans in green, South Asians in pink and mixed ancestry in blue. On the right, bar plot showing the percentage of each ethnicity.

3.3 Disease transmission

3.3.1 Sporadic cases

Multiple studies have reported that the prevalence of monogenic SRNS is higher in individuals with familial disease (60). Nevertheless, in absence of consanguinity or familial history, numerous genetic alterations have been found in sporadic cases specifically presenting autosomal dominant disorder with the identification of de novo mutations in disease causing genes. From the total 422 probands, 378 were affected by sporadic SRNS and did not have familial disease. Six patients had unknown family history and were presumed sporadic. Therefore, a majority of patients were sporadic cases representing 91% of the cohort. This demographic may have been influenced partly by some recruitment bias, since patients with mutations in the associated SRNS genes are often familial and the focus was on recruitment of cases where mutations had not been confirmed in some of the established genes.

3.3.1.1 Sporadic cases with extended family sequenced

Seven sporadic cases of the total 378 sporadic patients have at least one relative that underwent WES or WGS. WGS sporadic cases with extended relatives comprised of one trio and two duos. The trio included the affected daughter and her unaffected parents. One duo contained an affected mother and unaffected daughter and the other duo, an affected daughter with unaffected mother. Correspondingly, WES sporadic cases with extended relatives entailed one family with five members, two trios and one duo. The quintet encompassed an affected daughter, both unaffected parents and two unaffected brothers. One trio was formed by an affected daughter and two unaffected parents, whereas the other trio by an affected son and two unaffected parents. The duo consisted of an affected daughter and unaffected mother.

3.3.2 Familial cases – description of pedigrees

In ~9% of SRNS cases other family members have been affected by SRNS or another form of nephrotic syndrome. Within the cohort are 38 familial cases out of 422 cases. The familial cases can be divided into 14 probands and 24 singletons (families where only one member was sequenced) with additional family members sequenced, in total 24 affected family members and 32 unaffected relatives. Of the fourteen sequenced families; seven underwent whole exome sequencing, seven families underwent whole genome sequencing and three of those families have some members sequenced in both platforms (Table 6).

Table 6. Families of the cohort. The predigree size of each family with the total number of sequenced samples including affected and unaffected. Families that underwent WES were labelled with letters whereas families that underwent WGS were labelled with numbers.

Family	**Pedigree**	**Sequenced Samples**			**Sequencing**
ID	**Size**	Total	Affected	Unaffected	**Platform**
A	33	12	7	5	WES, 3xWGS
B	7	6	3	3	WES, 4xWGS
C	5	3	3	0	WES
D	5	3	3	0	WES
E	9	3	3	0	WES
F	4	2	2	0	WES
G	4	2	2	0	WES
1	5	5	2	3	WGS
2	5	5	2	3	WGS
3	7	3	2	1	WGS
4	5	3	2	1	WGS
5	4	3	2	1	WGS, 2xWES
6	3	2	2	0	WGS
7	3	2	2	0	WGS

3.3.2.1 WES Families

Detailed clinical evaluation including family history and pedigree was performed on each case and information from clinical notes as well as rare disease database entries was combined. Of the seven families that underwent WES, there were 1 family with twelve members, 1 family with six members, 3 trios and 2 duos. Apparent segregation patterns indicated that these comprise four autosomal dominant families and two with an autosomal recessive inheritance (Figure 21).

Family A was the largest studied, with seven affected and five unaffected members. Pedigree analysis suggested that SRNS is inherited as an autosomal dominant trait. Detailed clinical and family history was available on four generations, and members of three generations were available for sequencing. Of the seven affected relatives all had FSGS (4 renal biopsy proven). Four developed end stage renal failure, with one dialysis dependent and three post-transplant (no recurrence). One had CKD and proteinuria without progression to end stage renal failure despite being in his mid-70's, whereas the remaining two had mild proteinuria. Age of onset ranged from 8 to ~30 years of age. Therefore, there was a large variation in disease severity.

Family B also spanned three generations and had three affected members and three unaffected, with inheritance suggestive of autosomal dominant model. Of the three affected members, all had FSGS, however the proband only had mild proteinuria and normal renal function whereas the other two CKD, suggestive of incomplete penetrance. Here, the age of onset ranged from 8 to 35.

Family C had three affected siblings (two females and a male) and unaffected parents but were not sequenced. Consanguinity was confirmed after clinical review with parents being second cousins. Segregation of the disease in the pedigree was consistent

with autosomal recessive inheritance. Affected children have CKD and FSGS. The age of onset examined ranged from 6 to 14. Additionally, they have comorbidities such as microcephaly and learning difficulties suggestive of a genetic syndrome with renal involvement.

Family D consisted of three affected siblings (two females and a male). DNA was not available on the unaffected parents. Of the three siblings, one female had SSNS (and presumed MCD) with normal renal function. In contrast, the other two siblings had SRNS with FSGS on renal biopsy, progressed to end-stage renal disease. All had disease onset in early life, ranging from 11 months to 2 years. Thus, disease phenotype between siblings is very different.

Family E had two affected sisters and affected mother who were sequenced. The proband had secondary multidrug resistant SRNS and histological diagnosis of FSGS that responded to rituximab. In contrast, her mother and sister were diagnosed with SSNS and did respond to conventional treatment with steroid and MMF. Disease onset ranged from 3 to 6 years. The father and other two sisters were unaffected but unfortunately, DNA was not available.

Family F had two affected brothers and unaffected parents that were not sequenced. Both brothers were diagnosed with SRNS and FSGS and one had severe learning difficulties. Their age of onset ranged from 5 to 6.

Finally, Family G had two affected sisters and unaffected parents that were not sequenced. Both sisters were diagnosed with primary SRNS and FSGS and developed multidrug resistance, with the age of onset of disease ranging from 1 to 8 years. DNA was unavailable on parents.

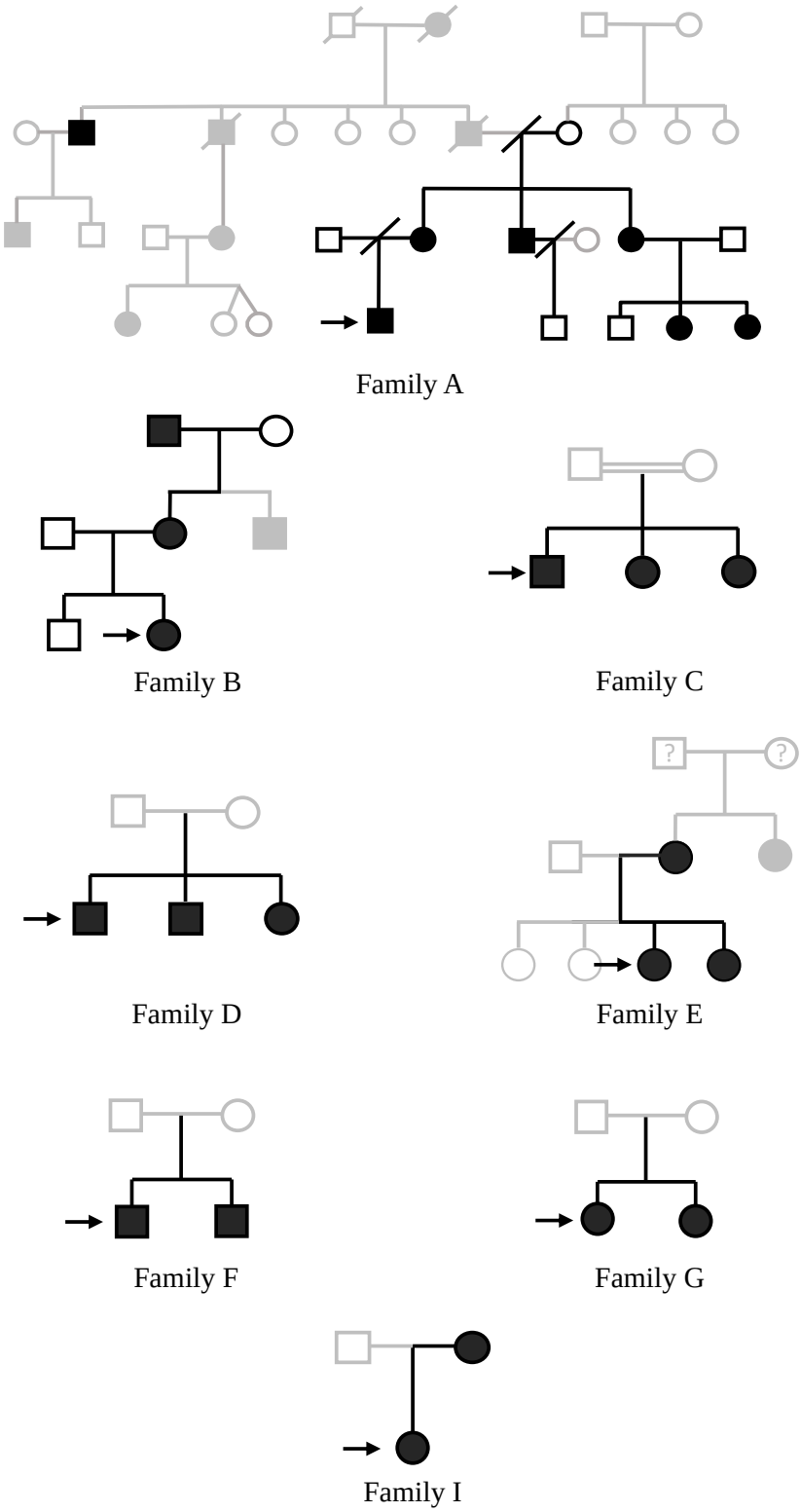


Figure 21. Pedigree structure of the families that underwent WES. Pedigree structure of seven families. Squares and circles indicate males and females respectively. Individuals coloured in black have been sequenced whereas individuals coloured in grey have not.

3.3.2.2 WGS Families

Of the seven families that were whole genome sequenced, there were 2 families with five members, 3 trios and 2 duos. Preliminary data suggests a family with an autosomal dominant model of inheritance (family 3) and another family with an autosomal recessive model (family 5) (Figure 22).

Family 1 has two affected members (brothers) and three unaffected (parents and daughter). Both brothers were diagnosed with SSNS with steroid dependence and MCD with childhood onset at 3 and 8 years. One recovered aged 12 years, with no further relapses whereas the other became steroid dependent requiring steroid sparing medications and did not go into remission.

Family 2 comprised IVF triplets (two affected males and an unaffected female) and unaffected parents. After examination of the sequence data, the triplets were confirmed non-identical and that the father was not the biological father (IVF was through an unrelated sperm donor). The two affected triplets developed childhood onset secondary SRNS at 2 years of age, responded to rituximab but remain medication dependent at 10 years.

Family 3 has two affected relatives (mother and daughter) and one unaffected (son). This family have two pregnancies not carried to term which suggest an autosomal dominant model of inheritance. Mother and daughter were diagnosed with SSNS and MCD; both had secondary SRNS with the mother responding to secondary agents and the daughter to rituximab.

Family 4 consisted of two affected half-sisters (different fathers) and unaffected mother. Both sisters were diagnosed with secondary SRNS, although one had FSGS

on kidney biopsy whereas the other MCD. One made a partial response and the other a complete response to rituximab.

Family 5 had two affected siblings and unaffected mother. Parents were first cousins. Both siblings had primary SRNS with an age of onset ranging between 7 to 10 years and chronic kidney disease.

Family 6 comprised two affected relatives, son and father, with primary SRNS and FSGS on renal biopsy as well as CKD. Additionally, the father had hypertension and the son, sickle cell disease.

Family 7 consisted of a son and mother, both with SSNS and MCD on biopsy. While both developed steroid dependence, they responded to secondary agents and mother was in permanent remission since her teens. Parents were first cousins, but other siblings were unaffected.

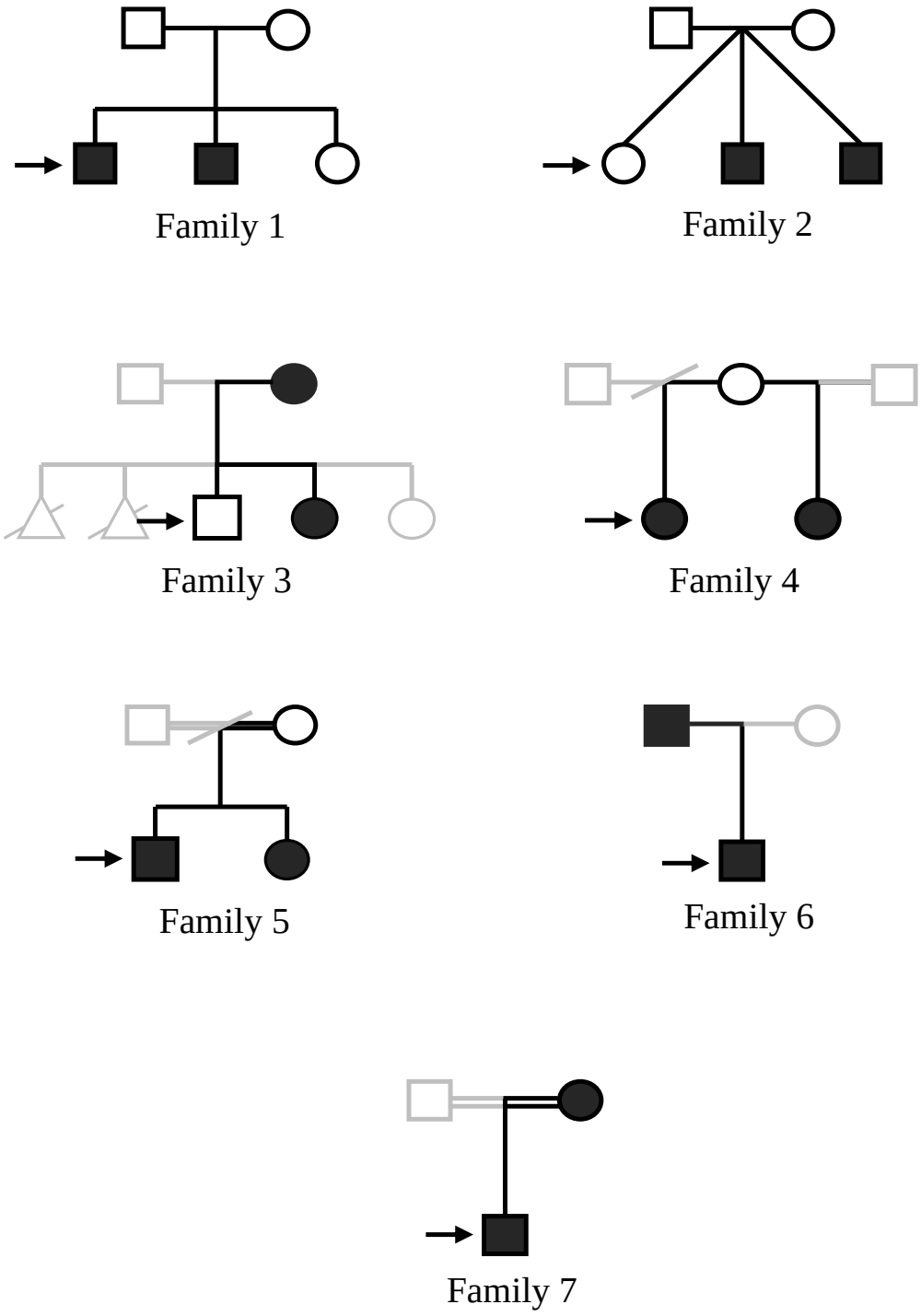


Figure 22. Pedigree structure of the families that underwent WGS. Pedigree structure of seven families. Squares and circles indicate males and females respectively. Triangles were used for pregnancies not carried to term. Individuals coloured in black have been sequenced whereas individuals coloured in grey have not.

3.4 Phenotyping of cases

Each patient was deeply phenotyped using detailed clinical information extracted from patient records and if needed, from sequential rare disease database entries from the UK Renal Disease Registry with the help of Dr. Ania Koziell. This ensured all information was contemporaneous and accurate reflecting any changes in clinical status or revisions in diagnosis. Additionally, it allowed for exclusion of cases with overlap pathology such as IgA nephropathy and secondary nephrotic syndromes misdiagnosed as primary SRNS. Follow up for selected patients ranged from 5 to 50 years. Database entries included parameters such as uploaded clinical records, renal biopsy, and results of routine haematology and blood chemistry such as serum creatinine and albumin as well as measure of urinary protein-creatinine ratios. A number of clinical features were annotated as follows: primary diagnosis, family history, age of onset, biopsy findings, nephrotic syndrome type, treatment, comorbidities, from sequential (anonymised) patient records that incorporated long-term clinical outcomes including which stage of chronic kidney disease was reached if any, whether the case developed end stage kidney failure and a need for renal replacement therapy and/or transplantation, and whether cases developed disease recurrence post-transplant. These clinical features were informative and provided a better understanding of the disease, but they were not independent variables as some of them were highly correlated.

The phenotype characteristics were annotated to build up a precise picture of the study cohort and to ensure deep phenotyping of all cases. This was crucial as the number of cases generally available for study in a rare disease are limited and can ultimately restrict the value of experimental approaches designed such as association studies especially if rigorous phenotyping is not performed.

3.4.1 Age of onset

The overall mean age of onset for the cohort was 15.4 (ranged from birth to 82). The majority of cases could be assigned an exact age of onset, aside from a percentage of adults that had childhood onset disease and did not recollect when this commenced. 366 patients out of 422 have the age of onset annotation. Cases were annotated as paediatric onset (PO) or adult onset (AO): 315 of 422 cases had PO, 97 had AO and 10 did not have information about onset available. Since SRNS is classically thought of as a childhood disease, surprisingly onset of SRNS appeared to follow a bimodal distribution, characterised by a first peak extending from birth to the second decade and a second peak from the fifth to sixth decades (Figure 23).

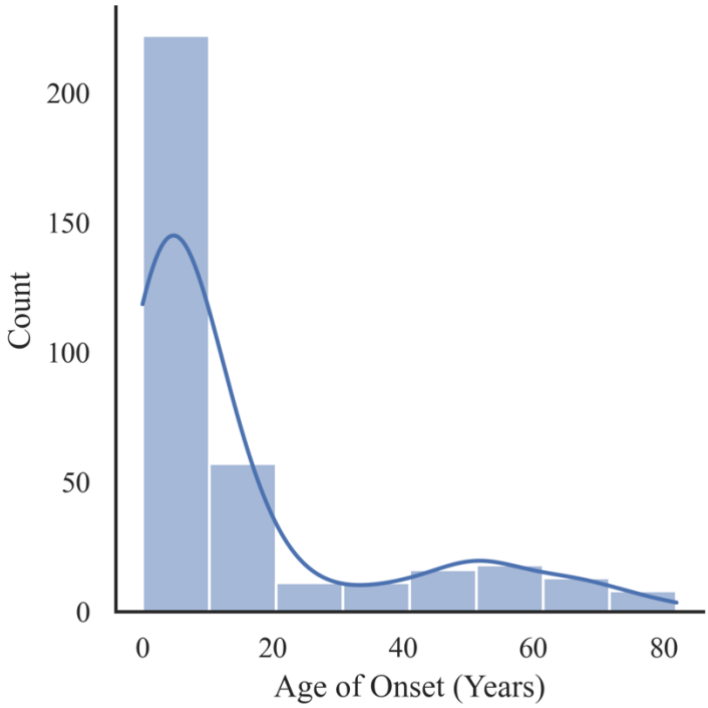


Figure 23. Age of onset distribution in SRNS patients (n=366). Empirically derived age of onset in years with admixture modelling.

Age of onset was also explored in relation to family history and the genetic findings described in Chapter 4. Cases with family history and likely Mendelian inheritance were found to have an earlier age of onset of 6.6 years (Figure 24). Whilst familial SRNS can occur in later life (e.g. *TRPC6* or *ACTN4* mediated disease), it is more prevalent in early life and this finding was consistent with previous studies (60, 82). Furthermore, patients with monogenic forms of SRNS that could be attributed to a causal genetic mutation in the established genes had the earliest age of onset of 2.9 years (Figure 24). Therefore, there was an overlap between the cases with family history and cases with monogenic forms of the disease. Some of these cases with a family history had adult onset whereas cases where a causal genetic mutation was identify only had paediatric onset. Cases with genetic risk factors for *APOL1* (G1 or G2 allele) were not included in Figure 24.

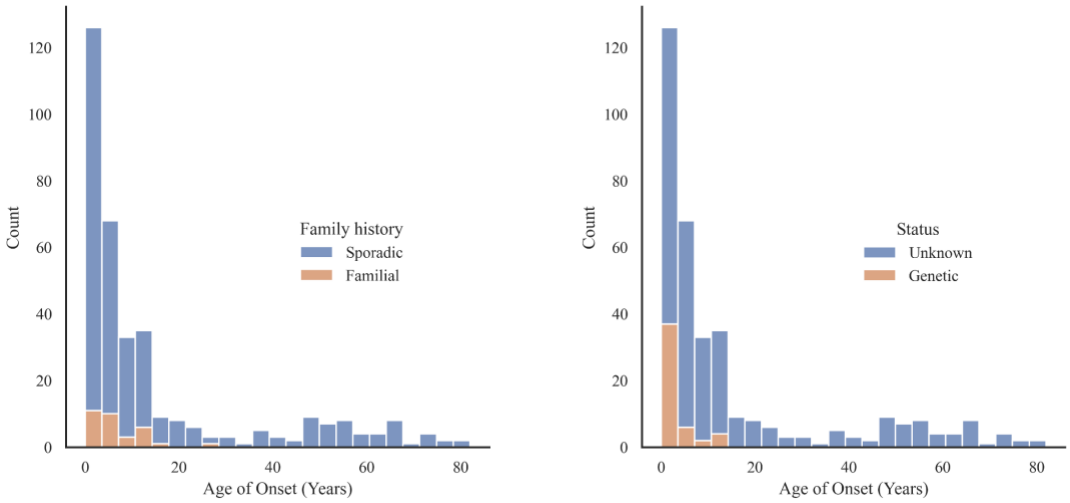


Figure 24. Age of onset of SRNS in patients with family history and patients with monogenic disease. On the left, the distribution of the age of onset for familial cases (orange) vs sporadic cases (blue). On the right, the distribution of the age of onset for patients with a confirmed monogenic disease (orange) vs patient with unknown genetic cause (blue).

3.4.2 Nephrotic syndrome type

Steroid response is used to describe the type of nephrotic syndrome as steroid sensitive or steroid resistant. This remains the working classification applicable clinically in the absence of alternative biomarkers for disease stratification. In this study, classical definitions were used to define cases as either steroid sensitive (SSNS) – response within one month for children or four months for adults, or steroid resistant (SRNS) if there was no response within one month for children or four months for adults. SRNS was subdivided into primary SRNS, cases that showed no response to steroid even at the outset, and were also often multidrug resistant and secondary SRNS, those patients that subsequently developed steroid resistance after initial response. Of the total 422 individuals, 253 (60%) were diagnosed with primary SRNS, 154 (37%) were diagnosed with secondary SRNS, 9 (2%) were diagnosed with SSNS and 6 (1%) did not have a diagnostic tag.

3.4.3 Histology

The most common renal histology in primary SRNS was FSGS, and MCD for secondary SRNS. Biopsy findings indicated FSGS in 253 patients, with a minority of syndromic childhood SRNS (specifically Pierson, Denys Drash and Frasier syndromes) showing a mixed picture of FSGS combined with diffuse mesangial sclerosis. Two cases with FSGS associated with mild GBM abnormalities and *de novo* COL4A mutations suggestive of Alport syndrome but had renal restricted phenotype and no family history. MCD (including its variants IgM and C1q nephropathy) was detected in 143 patients, histological changes compatible with congenital nephrotic syndrome in 13 patients and no histology annotation was available for 13 patients.

3.4.4 Clinical outcome

Primary SRNS with FSGS as a histological pattern are associated with a higher risk of developing end-stage kidney disease and responsible for 15% of all children with CKD (39). As seen within the clinical demographic of this cohort, primary SRNS patients may progress to ESRF more rapidly if a genetic cause is detected and/or the FSGS is part of a syndrome. Others responded to treatment and may achieve partial remission or complete recovery, especially if the SRNS is secondary.

Patients in this cohort were grouped according to the following clinical outcomes: recovery, stable on medication and CKD/ESRD/Transplant. Recovery was defined as no relapse for 5 years off medication. The group stable on treatment (with or without secondary agent) encompassed patients in remission (urine testing negative for protein) or with mild protein loss. The final group included those cases that developed multidrug resistance, progressing to CKD or ESRD and dialysis or renal transplant, and whether post-transplant recurrence occurred. Among 315 patients with information about their clinical outcome, 42 achieved recovery, 128 were stable on medication and 145 patients progressed to ESRD/CKD or transplant. 27 patients of these 145 developed post-transplant disease recurrences. To summarise, despite clinical outcomes for SRNS appearing heterogeneous, the majority of cases had ESRD/CKD or were stable on medication.

3.5 Discussion

This chapter comprehensively explored the recruitment and characterisation of an SRNS cohort (primary and secondary) of children and adults from UK nephrology centres. To my knowledge these patients represent the largest SRNS cohort that has undergone whole exome and/or whole genome sequencing in the UK and one of the

most extensive resources that includes genetic data and phenotype information internationally, with the exception of the study performed by Hildebrandt's group that included 1783 families (82). Phenotyping involved extraction of a detailed clinical course, including parameters such as age of onset, outcomes, biopsy data and medication history using clinical records and Radar/Renal Registry to allow rigorous definition of the phenotype. Accessing the actual phenotype within a nationally recruited cohort can be particularly challenging – either through inaccessibility of data, or incomplete or inaccurate database entries since this may not be performed by specialists in the field. Another disadvantage is that records may not be updated over time as cases move between paediatric and adult care, or hospital trusts and the thread describing clinical progression is lost.

Sporadic SRNS in the European population was accurately represented within this cohort since this was the main group recruited. The exploration of the disease transmission showed 91% of the patients had sporadic SRNS and only 9% had family inheritance. The clinical characteristics available for these patients highlight the heterogeneous nature of the disease and the variety of phenotypes. The results described in this chapter regarding sex ratio, age of onset in familial cases and clinical outcomes were consistent with previous studies of SRNS in children and adults (38, 39, 60, 82). The bimodal distribution found in the age of onset was a novel finding as classically, this is considered a disease of early life. Nevertheless, the epidemiological observations made from this cohort are still limited due to a lack of statistical power consequent on an insufficient number of cases because of the rare status of the disease.

The male to female ratio within this cohort was 1.19:1, which is broadly consistent with other studies (82, 159) although the differences are not as striking. The reduced

ratio is likely related to the demographic of the cohort as there was an almost equal collection of children and adults, and majority of SRNS studies only include children. In order to demonstrate that SRNS is more common in males than females, all new patients arriving in a clinic over a period of time would need to be catalogued. Furthermore, there are other sources of bias during the recruitment of genetic studies that could affect the sex ratio (160).

Fourteen families of different sizes that had detailed phenotypic information available were whole exome and/or whole genome sequenced. At least two families (family D and family E) presented with a mixed phenotype; some members presented with classical SRNS and others with SSNS. This mixture of phenotypes within the same family highlights the heterogeneity of INS and the similarities of SRNS and SSNS that could potentially have a similar or shared genetic architecture. Notably, even when all the affected members of a family had SRNS, there was still phenotypic variability in phenotypic expression. This could be due to incomplete penetrance or epistatic effects such as modifier genes.

The main objective of treating INS is to achieve long-term nephrotic remission whilst minimise toxic side effects of medications and complications such as hypovolaemia, infection and thrombosis. Assessing the clinical outcomes of SRNS based on response to treatment, genetics and clinical features such as histology or age of onset can be challenging because of the limited size of the cohorts available for study, reduced follow-up period and lack of information for some patients with unknown or uncertain features. Furthermore, despite national standardisation, clinics do not always adopt national guidelines and may have different protocols meaning that SRNS patients are not offered a standard set of medications. This can also be due to funding issues and

other factors, for example Rituximab is used for complex SRNS as protocol, especially for children with secondary steroid resistance but this is not as prevalent in adult nephrology practice (22, 36). Thus, it is difficult to standardise clinical outcomes, since observed differences in clinical features between patients could be linked to the consequences of how patients are managed and treated, rather than their underlying genetics making stratification on this basis challenging. As the majority of SRNS cases had an early age of onset (< 18 years) and were diagnosed with primary SRNS that is classically multidrug resistant, the most frequent clinical outcome was CKD, ESRD or transplant, followed by patients stabilised on multiple medications but not in permanent disease remission. The outcomes identified were consisted with those described in previous clinical studies, which provided further verification that despite being a relatively small cohort that did not capture all SRNS cases in the UK, the study did accurately represent the SRNS phenotype.

.

Chapter 4 – Evaluation of rare genetic variants disrupting coding regions of established SRNS genes

4.1 Study design considerations

Despite the direct association of mutations in 67 genes with SRNS, the molecular basis of the disease remains poorly understood as only a maximum of 29.5% of patients (with familial or sporadic disease) carry a causal mutation in one of these established genes (20, 82). Furthermore, patients present considerable variability in their phenotype, even in families where affected relatives share the same causative mutation. This supports the hypothesis that as well as monogenic causes, modifier genes, polygenic risks and complex environmental-genetic interactions could potentially modulate disease severity. These mechanisms remain largely unexplored in SRNS.

Within this study, even if a candidate gene had been reported as causing SRNS or comparable disease phenotype, the evidence for this from previous studies was carefully evaluated. Thus, in order to identify disease-causing or associated variants, it is crucial to separate genuine causal variants from the background of variants that are rare, but not actually pathogenic for SRNS. A pathogenic variant significantly alters levels or affects the normal biochemical function of the product of the affected gene, playing a role in the disease pathogenesis. In this study, certain parameters of pathogenicity were set; synonymous variants were not considered as putative pathogenic variants since they are not expected to have a damaging effect on the protein structure, despite the occasional association with splice site formation. However, rare variants located within coding regions of established SRNS genes and

intronic variants adjacent to exons (±10 bp) that can cause abnormal splicing by interfering with splice site recognition were examined. Predicted distributions of allele frequencies and effect sizes for pathogenic variants were also taken into consideration, to determine the probability of observing this result by chance with a randomly selected variant. If a variant is pathogenic, it will be generally subject to negative selection and be rare in a healthy population. How rare also depends on its penetrance and the mode by which it is known to cause disease (118). Since nephrotic syndrome is a rare disease, variants with an allele frequency higher than 0.001 were excluded as these were highly unlikely to be pathogenic. However, some common variants known to be predisposing factors for SRNS were also explored separately, such as the risk alleles G1 and G2 in *APOL1* (84) and the *NPHS2* gene mutation R229Q encoding an allele associated with albuminuria and ESRD when inherited in *trans* with another *NPHS2* mutation (75). In addition, the potential pathogenicity of a variant was assessed by deleteriousness rankings from CADD algorithm (161). CADD score was used to predict how damaging a variant might be on protein function, based on conservation across species and difference between the reference and alternative amino acid. Therefore, variants with a score higher than 15 were prioritised in the first instance. Data from each patient was analysed individually, assuming either a dominant or recessive model based on the established inheritance of the gene.

Sequences from 422 probands of European, South Asian, East Asian, African and mixed-race descent (Figure 20), were analysed to confirm the presence of potential mutations in the known SRNS genes. Rare and functional variants in the established 67 genes were extracted in each sample. This included variants with an allele frequency lower than 0.001 for dominant model and 0.01 for recessive alleles and variants that affected the protein coding sequence such as nonsense, frameshift, small

in-frame insertions or deletions, splice acceptor and stop gained or lost (section 2.5.4). In cases where a potentially pathogenic mutation was detected, available family members were also investigated to determine if the mutation was segregating with the affected members of the family. Finally, the assessment of structural variation, such as copy number variation (CNV), across the established SRNS genes was beyond the scope of this thesis.

4.2 Results

Potentially pathogenic variants in established SRNS genes were classified into three main categories: variants that have been previously described in the literature to be responsible for SRNS (Table 6), variants that are novel or have not been reported in the literature (Table 7) and variants previously described or novel, that had an inconsistent model of inheritance to the established gene (Table 8). The latter are variants that have been reported causative in homozygous and were found in heterozygous in the cohort (Table 8).

A total of 67 pathogenic or likely damaging variants were detected across 48 cases of the 422 (Table 7) (Table 8). These harboured either pathogenic or likely pathogenic variants with genotypes that were consistent with the described mode of inheritance of the gene in question and robustly explained the associated phenotype based on existing data from previous studies (Table 7) (Table 8). Of these, 19 samples were investigated for compound heterozygosity as they had two potentially causal variants identified within the same gene. When possible, any compound heterozygous mutations were confirmed by sequencing the parents of the case. Out of the 48 patients where mutations were identified, 11 (23%) had familial disease and 4 of them had relatives sequenced in the cohort whereas 38 (77%) patients were sporadic cases with

no apparent family history of kidney disease. The 4 probands and family members were investigated and a clear link between their genotype and phenotype was identified as mutations were present in affected members and absent in unaffected members.

All affected individuals with a potential mutation identified had paediatric onset, 47 were diagnosed with primary SRNS and 1 with secondary SRNS, 23 were female and 25 were male. Their ethnicity can be summarised as followed: 33 Europeans, 13 South Asians, 1 East Asians and 1 African.

All potential pathogenic variants were found in coding regions except for seven variants that were close to the exons and altered splicing. Only 13 genes out of the 67 that have been directly associated with SRNS, where found to harbour mutations in this cohort (Figure 25.A). The genes harbouring the biggest number of mutations within the cohort were *NPHS1* and *NPHS2* (Figure 25.A), replicating the results of some other studies and validating the cohort to some extent as typical of SRNS and therefore representative (60, 80). The type of mutation of these variants and consequences at the protein level is showed in Figure 25.B, with nonsynonymous single nucleotide variants being the most frequently observed, followed by splicing, indels (frameshift and non-frameshift), stop gained variants and nonframeshift deletion. Information about the reported status can be also found in Figure 25.B.

The distribution of the allele frequency of the mutations was calculated using gnomAD (133) and divided into two groups, heterozygous and homozygous variants, showing a range of 0 to 0.00125 for heterozygous variants and a range of 0 to 0.0006 for homozygous variants. A heatmap plot showing the breakdown of mutations by type of substitution reveals a high frequency of C>T and G>A mutations (Figure 25.C).

Additionally, affected individuals were interrogated for *APOL1* G1 allele and G2 alleles, known to be risk alleles associated with FSGS, CKD/ESRF and hypertension in the African population (86, 162). Of the total 422 cases in the study cohort, only 35 had African ancestry and from those, 23 presented at least one risk allele in the G1 allele and/or G2 allele, known to be associated with pathogenicity and an increased risk of developing SRNS (Table 10). Additionally, these same risk alleles were detected in 6 cases that did not cluster closely with individuals of African ancestry on PCA plot but were known mixed race with African influence.

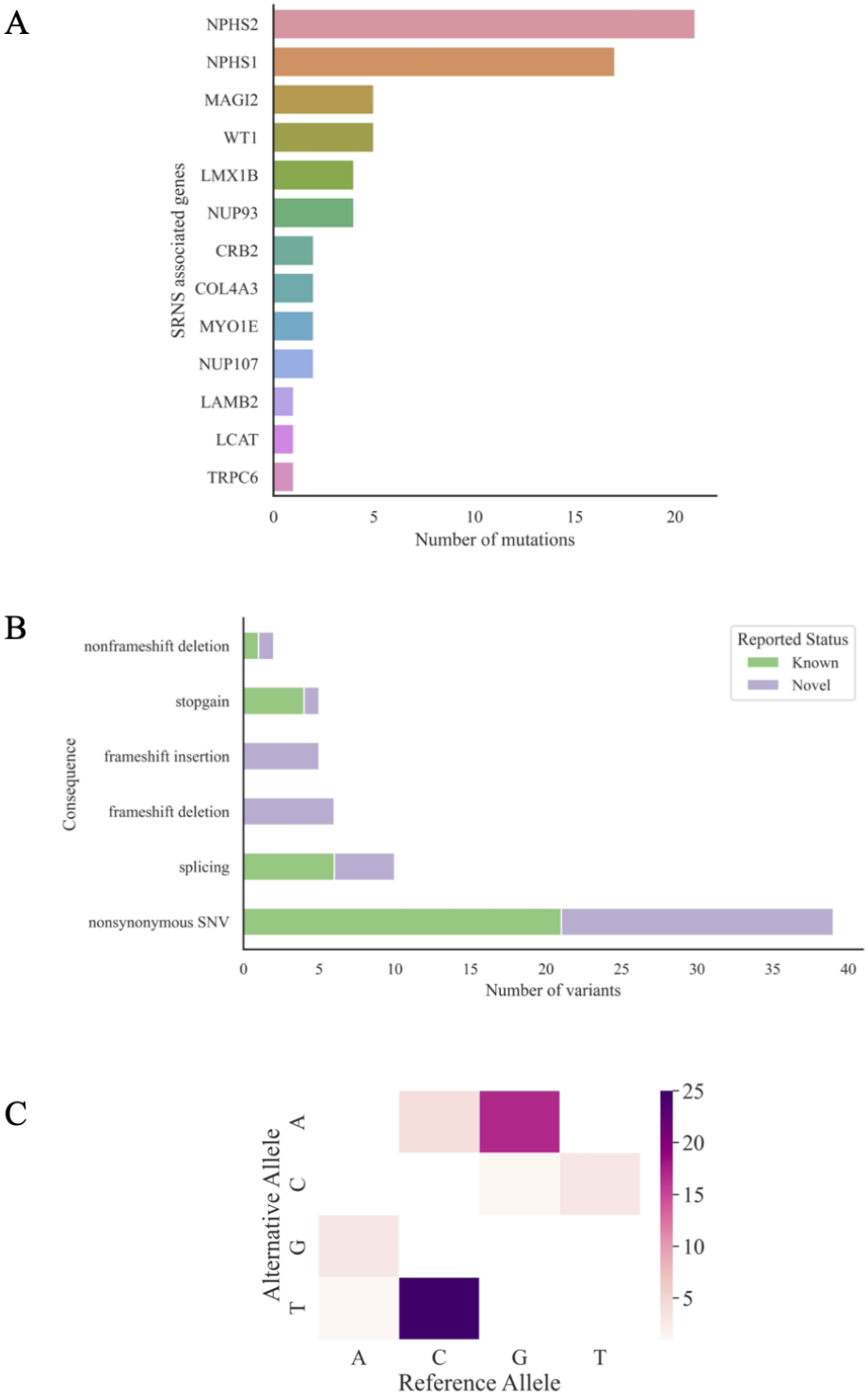


Figure 25. Analysis of known causative variants identified in the cohort.(A) The breakdown of associated SRNS genes where mutations were found in the cohort. 67 pathogenic variants were detected across 48 patients (B) Distribution of consequence classes and the reported status of the variants: known variants presented in green and novel in purple. (C) Heatmap plot showing the breakdown of mutations by type of substitution.

Table 7. Previously described variants in established SRNS genes. The genome build used was GRCh37.

Sample	Gene	Zygosity	Chr	Pos	Ref	Alt	Consequence
S1228	*LAMB2*	HOM	chr3	49168562	G	A	p.R246W
S2227	*LMX1B*	HET	chr9	129455598	G	A	p.R246Q
S012538	*LMX1B*	HET	chr9	129455598	G	C	p.R246P
S1852	*LMX1B*	HET	chr9	129455598	G	A	p.R246Q
S1217	*NPHS1*	HOM	chr19	36321958	G	A	p.R1160X
S1223	*NPHS1*	HET	chr19	36333453	C	T	splicing
	NPHS1	HET	chr19	36333296	G	A	p.R831C
S1654	*NPHS1*	HOM	chr19	36341889	G	A	p.P167L
S1704	*NPHS1*	HET	chr19	36342241	G	A	p.A107V
	NPHS1	HET	chr19	36342681	C	A	splicing
S1851	*NPHS1*	HET	chr19	36333453	C	T	splicing
	NPHS1	HET	chr19	36321994	G	A	p.Q1148X
S0635	*NPHS2*	HOM	chr1	179530462	C	T	p.R138Q
S0694	*NPHS2*	HOM	chr1	179526187	CTCTC	-	p.235_238del
S1214	*NPHS2*	HET	chr1	179530462	C	T	p.R138Q
	NPHS2	HET	chr1	179526257	G	A	p.Q215X
S1652	*NPHS2*	HOM	chr1	179530462	C	T	p.R138Q
S1668	*NPHS2*	HET	chr1	179530462	C	T	p.R138Q
	NPHS2	HET	chr1	179526257	G	A	p.Q215X
S1714	*NPHS2*	HET	chr1	179526214	C	T	p.R229Q
	NPHS2	HET	chr1	179521740	G	A	p.R223W
S1726	*NPHS2*	HET	chr1	179526214	C	T	p.R229Q
	NPHS2	HET	chr1	179521740	G	A	p.R223W
S1731	*NPHS2*	HET	chr1	179530462	C	T	p.R138Q
	NPHS2	HET	chr1	179521743	C	T	p.V222M
S0621	*NUP107*	HOM	chr12	69115634	G	A	p.C442Y
S1696	*NUP107*	HOM	chr12	69084526	G	A	p.M101I
S0689	*WT1*	HET	chr11	32414250	C	T	p.R439H
S1211	*WT1*	HET	chr11	32413566	G	A	p.R467W
S1677	*WT1*	HET	chr11	32413513	C	T	splicing
S1850	*WT1*	HET	chr11	32413513	C	T	splicing
S014195	*WT1*	HET	chr11	32413514	G	A	splicing

Table 8. Novel variants in established SRNS genes. The genome build used was GRCh37.

Sample	Gene	Zygosity	Chr	Pos	Ref	Alt	Consequence
S1848	*COL4A3*	HET	chr2	228147214	A	-	p.G874fs
	COL4A3	HET	chr2	228175539	T	-	p.P1601fs
S2032	*CRB2*	HET	chr9	126133214	C	T	p.R628C
	CRB2	HET	chr9	126135914	-	GGCCC	p.P1035fs
S2054	*LCAT*	HOM	chr16	67977109	C	A	p.G54V
S1684	*LMX1B*	HET	chr9	129455537	C	T	p.L226F
S1722	*MAGI2*	HOM	chr7	77649003	C	-	p.G1319fs
S1728	*MAGI2*	HET	chr7	77755055	-	GCCAGT	p.R1161fs
	MAGI2	HET	chr7	79082567	GGTTG	-	p.G21fs
S1717	*MAGI2*	HOM	chr7	77824247	C	T	p.R738Q
S1682	*MAGI2*	HOM	chr7	77824247	C	T	p.R738Q
S0727	*MYO1E*	HOM	chr15	59466395	A	T	p.Y698X
S1230	*MYO1E*	HOM	chr15	59528852	T	C	p.K118E
S1215	*NPHS1*	HET	chr19	36334481	G	A	p.R743C
	NPHS1	HET	chr19	36333400	C	T	p.G796E
S1655	*NPHS1*	HET	chr19	36333453	C	T	splicing
	NPHS1	HET	chr19	36330191	-	C	p.L1019fs
S1660	*NPHS1*	HOM	chr19	36330304	-	T	p.T982fs
S1669	*NPHS1*	HET	chr19	36321954	C	A	splicing
	NPHS1	HET	chr19	36322587	-	C	p.G1082fs
S1676	*NPHS1*	HOM	chr19	36336290	AAG	-	p.636_637del
S010304	*NPHS1*	HOM	chr19	36336272	A	G	p.L643P
S0646	*NPHS2*	HET	chr1	179530462	C	T	p.R138Q
	NPHS2	HET	chr1	179533820	C	T	splicing
S1673	*NPHS2*	HET	chr1	179530462	C	T	p.R138Q
	NPHS2	HET	chr1	179533824	C	A	splicing
S1715	*NPHS2*	HET	chr1	179530462	C	T	p.R138Q
	NPHS2	HET	chr1	179544847	C	-	p.A51fs
S014334	*NPHS2*	HET	chr1	179530462	C	T	p.R138Q
	NPHS2	HET	chr1	179521756	TT	-	p.Q217fs
S0648	*NUP93*	HET	chr16	56875663	T	C	p.L633S
	NUP93	HET	chr16	56872929	T	C	p.L572S
S2960	*NUP93*	HOM	chr16	56871529	A	G	p.K514E
S013682	*NUP93*	HOM	chr16	56871529	A	G	p.K514E
S1140	*TRPC6*	HET	chr11	101323814	C	T	p.D890N

Table 9. Previously described or novel variants in established genes with an inconsistent model of inheritance. The genome build used was GRCh37

Sample	Gene	Zygosity	Chr	Pos	Ref	Alt	Consequence
S1656	*COL4A1*	HET	chr13	110815866	G	A	p.P1398L
S0628	*NPHS1*	HET	chr19	36322018	G	A	p.R1140C
S0629	*NPHS1*	HET	chr19	36342697	C	G	p.G15R
S1656	*NPHS1*	HET	chr19	36340184	C	G	p.C265S
S1689	*NPHS1*	HET	chr19	36342248	C	T	p.D105N
S2051	*NPHS1*	HET	chr19	36321796	T	C	p.T1182A
S2232	*NPHS1*	HET	chr19	36317544	G	A	p.P1200S
S009997	*NPHS1*	HET	chr19	36321820	G	A	p.H1174Y
S010112	*NPHS1*	HET	chr19	36339613	T	G	p.S366R
S013742	*NPHS1*	HET	chr19	36336398	C	G	p.G601A
S014148	*NPHS1*	HET	chr19	36340506	A	C	p.S220A
S014207	*NPHS1*	HET	chr19	36317420	A	G	p.V1241A
S014249	*NPHS1*	HET	chr19	36333453	C	T	splicing
S014285	*NPHS1*	HET	chr19	36332686	C	A	p.A916S
S1678	*NPHS1*	HET	chr19	36321958	G	A	p.R1160X
S0760	*NPHS1*	HET	chr19	36336953	G	A	p.C528C
	NPHS1	HET	chr19	36342236	A	G	p.Y109H
S014415	*NPHS1*	HET	chr19	36333024	A	T	splicing
S1686	*NPHS2*	HET	chr1	179533907	C	T	p.G99E
S1846	*NPHS2*	HET	chr1	179520568	C	G	p.E230Q
S2053	*NPHS2*	HET	chr1	179530421	-	T	splicing
S2231	*NPHS2*	HET	chr1	179528830	T	C	p.E173G

Table 10. Number of *APOL1* risk alleles in cases. The three APOL1 alleles were:G1,G2 and WT (wild-type).

Genotype	Cases
WT+G1	10
WT+G2	4
G1+G1	13
G2+G2	2
G1+G2	2

4.2.1 Previously described variants in established genes

A total of 24 cases had at least one putative pathogenic variant in an associated SRNS gene that has been previously described in association with familial or sporadic SRNS (Table 6). Visual inspection of the read alignments was performed by IGV (Integrative Genomic Viewer) (163) in order to verify the 32 variants in each gene of interest. All these variants were well characterised and had been confirmed pathogenic in previous studies. This established a clear link between genotype and phenotype in these 24 cases and therefore, were subsequently considered 'solved' and not included in any further analysis. Additionally, each mutation was confirmed by Sanger sequencing by Dr. Agnieska Bierzynska and Dr. Ania Koziell (60, 102) (Table 6).

An interesting finding was that some mutations occurring in the *NPHS2, LMX1B* and *WT1* genes were identified in multiple unrelated cases. This suggested that mutations within these genes are prevalent across sporadic SRNS cases in early life as well as familial. The *NPHS2* gene mutation p.R138Q (rs74315342) was detected in five unrelated sporadic cases within the cohort. This mutation has previously been associated with familial SRNS occurring in early life (164-166). Four of these cases (S0635, S1214, S1652 and S1668) had a very early onset (<1 year), whereas S1731 had a later onset (13 years). The *LMX1B* mutation p.R246Q (rs1191455921) was also detected in two patients, one sporadic (S2227) and another familial (S1852) though no relatives were available for sequencing to confirm this was inherited. Both cases had an early onset in childhood. The mutation p.R246Q has been previously described in Nail Patella syndrome and also in early onset FSGS (167, 168). Regarding *WT1* gene mutations, the splicing variant, rs587776576, was identified in S1677 and S1850 and has been associated with Frasier syndrome, SRNS and FSGS (Table 6).

Of the total 24 cases with previously described mutations in known SRNS genes, 5 had familial disease. S1696 was part of Family C, three affected siblings (two females and male). Here, a homozygous mutation in the *NUP107* gene (p.M101I) was detected in all three siblings. DNA was not available for sequencing on the unaffected parents, but consanguinity was confirmed through clinical review as parents reported that they were first cousins. All siblings presented with FSGS and CKD, as well as other comorbidities such as microcephaly and learning difficulties (Figure 26). This same *NUP107* gene mutation had been previously described in other families in association with both SRNS and intellectual disability, microcephaly and developmental delay (169, 170).

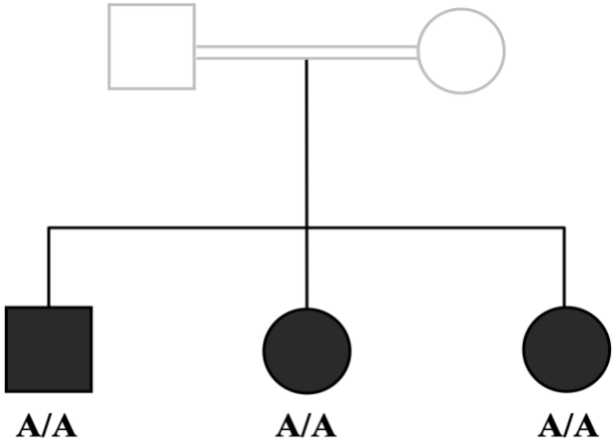


Figure 26. Segregation of the variant in *NUP107* (p.M101I) in Family C. Squares and circles indicate males and females respectively. All siblings were homozygous for the mutation. DNA was not available for sequencing on the unaffected parents. Consanguinity was confirmed through clinical review.

4.2.1.1 Risk alleles in APOL1

Data from studies carried out in the United States describes the risk of CKD and ESRF being 4 to 5 times higher in African Americans compared with individuals of European ancestry (171). Specifically, two particular risk alleles in *APOL1* have been associated with large increases in risk of kidney disease in both paediatric and adult patients of African descent. These risk alleles, annotated as G1 and G2, are associated with an increased risk of ESRD and FSGS by 7 to 10-fold (84, 86, 162).

Two mutations are responsible for the alteration of the amino acid sequence of the protein product and are located in the coding region of the C-terminal domain. The G1 risk allele results from two non-synonymous amino acid substitutions (rs73885319 and rs60910145) that are in linkage disequilibrium. The G2 risk allele is an in-frame deletion of two amino acid residues (rs71785313). Both alleles are present in 10-15% of individuals of African descent (172). The increase in renal disease risk is associated with any of the two alleles found under an autosomal recessive model of inheritance, although a smaller risk in individuals with one allele has also been described (162). Additionally, the combination of both alleles G1+G2 has also been associated with an increased risk. Since G1 and G2 are located within the same region in close proximity, recombination between them is very unlikely. Thus, G1 and G2 are mutually exclusive and the presence of both alleles in a single individuals indicated compound heterozygosity (84).

Although the majority of cases recruited to the study cohort were Europeans, 35 out of 422 cases had African descent. Of this subgroup of 35, at least one risk allele for G1 or G2 was detected in a majority of 25 cases (72%) whereas the remaining 10 cases (28%) did not carry any *APOL1* risk alleles (wild-type) (Figure 27.A). Therefore, more

than half of affected individuals of African descent in the cohort had a risk allele in *APOL1* (72%), this suggested a higher prevalence of G1 and G2 in SRNS cases compared with standard metrics in African populations of 10-15% (173). Moreover, the 6 cases that did not cluster with the individuals of African but were of mixed African/European origin also carried risk alleles in G1 or G2 (Figure 27.A). Thus, there were 31 SRNS cases with at least one *APOL1* risk allele in the cohort that did not have any known or likely pathogenic variants in established SRNS genes. The breakdown can be summarised as followed: 10 individuals were heterozygous for G1, 4 were heterozygous for G2, 13 were homozygous for G1, 2 were homozygous for G2 and finally, 2 samples had both risk alleles G1 and G2 risk alleles (compound heterozygosity) (Table 10).

The allelic inheritance involving the common genetic variants G1 and G2 in *APOL1* with small to moderate effects resulted in very heterogeneous phenotype across the cases. Affected individuals with monogenic form of SRNS where a genetic mutation with large effects on disease status was identified had early age of onset with a mean of 2.92 (Figure 24) whereas cases with a risk allele in APOL1 (G1 or G2) had a diverse age of onset that ranged from 1 to 65 years (mean=16.49) (Figure 27.B). Clinical outcomes for cases with a risk allele were also very heterogeneous. However, most cases had chronic kidney disease or required a kidney transplant.

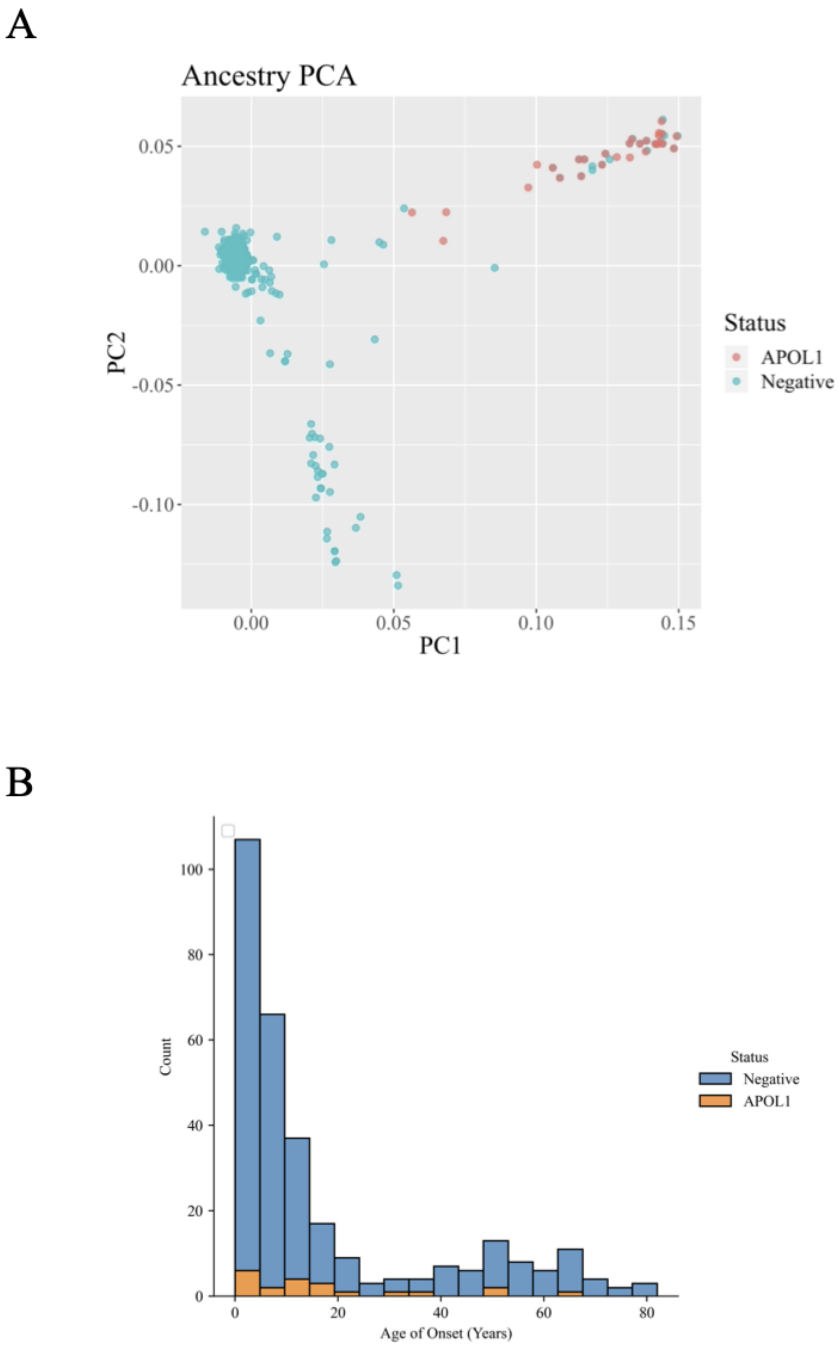


Figure 27. Principal component analysis of affected individuals with *APOL1* risks alleles and their age of onset. (A) Projection of individuals onto the first two principal components of genetic variation. Individuals with a risk allele in *APOL1* are represented in red whereas individuals with no risk allele are represented in blue. (B) Distribution of the age of onset for cases with a *APOL1* risk allele (orange) vs cases with no risk allele (blue).

Interestingly, Family 6 had two affected relatives (son and father) diagnosed with FSGS and CKD and who had the G1 risk allele. The father was heterozygous for G1 and had milder disease with a late age of onset at 30 years old, whereas the son was homozygous for G1 and had a more aggressive phenotype with an early age of onset at 9 years old and more rapidly progressing disease reaching CKD by 15 years of age. Both had other comorbidities that could potentially be associated with the presence of the G1 risk allele such as hypertension and sickle cell (162) (Figure 28).

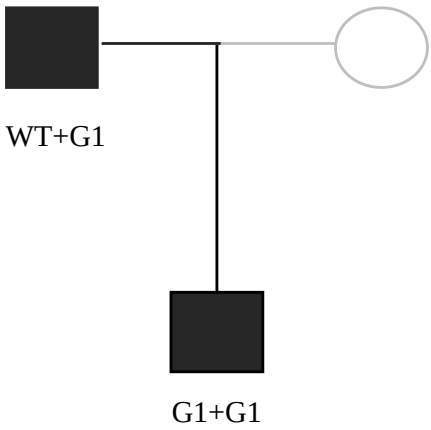


Figure 28. Pedigree structure of the Family 6 and their *APOL1* risk alleles. Squares and circles indicate males and females respectively. The risk alleles were shown as WT(wild type allele) and G1.

4.2.2 Novel variants in established genes

A total of 24 cases had at least one likely damaging variant that was not previously reported to be associated with SRNS or it was novel (never seen in another publicly available dataset) (Table 7). The genotype of these patients was consistent with the known mode of inheritance of the established SRNS gene in which the variant occurred. Furthermore, these 35 novels (or not reported) variants were predicted to be damaging to the encoded product according to different bioinformatic tools such as the ones described in the section 2.5.4.2. Some patients exhibited compound

heterozygosity between a previously described variant and a novel one (Table 8). Visual inspection of read alignments of each variant was performed with IGV. Moreover, these variants were also checked by Sanger sequencing.

Six cases with a novel variant in an established SRNS gene had familial disease, and three of them (S2054, S1673 and S1140) had relatives sequenced within the cohort. The three families were: Family 5, Family B and Family G.

Family 5 had two affected siblings with a novel homozygous variant in *LCAT* (p.G54V) and the unaffected mother was heterozygous for the variant. The father was not sequenced but parents of the affected siblings were first cousins. Both siblings were diagnosed with SRNS and CKD with an early age of onset (7 and 10 years) (Figure 29). *LCAT* is an enzyme involved in the cholesterol homeostasis and recently mutations in this gene have been linked to nephrotic syndrome and ESRD (174).

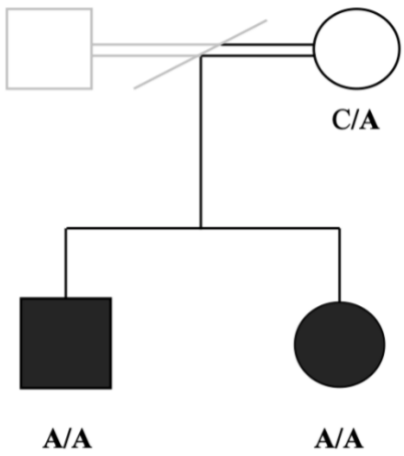


Figure 29. Segregation of the variant in *LCAT* (p.G54V) in Family 5. Squares and circles indicate males and females respectively. Both siblings were homozygous for the mutation whereas the mother was heterozygous. DNA was not available for sequencing on the unaffected father. Consanguinity was confirmed through clinical review.

Family B had three affected relatives (proband, mother and grandfather) where a novel heterozygous TRPC6 variant (p.D890N) was found. One had mild proteinuria and two had CKD and FSGS. The mutation was predicted to be pathogenic and was not present in the three unaffected members of the family (Figure 30).

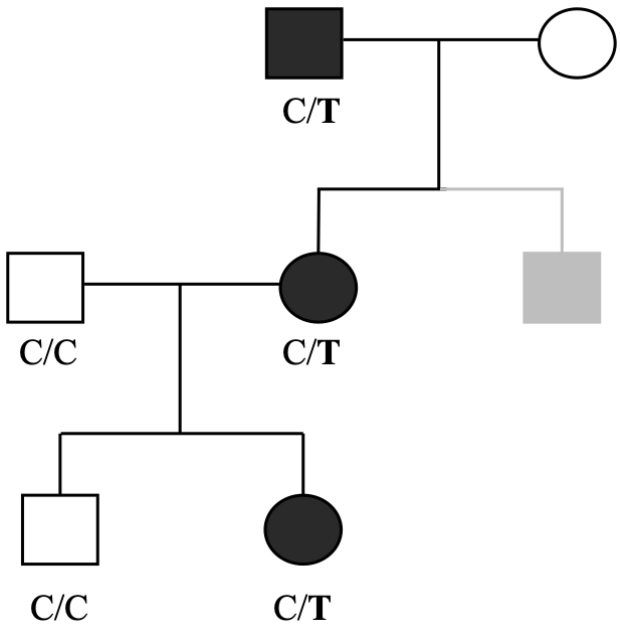


Figure 30. Segregation of the variant in *TRPC6* (p.D890N) in Family B. Squares and circles indicate males and females respectively. Family spanned three generations. Grandfather and mother had CKD disease whereas the proband only had mild proteinuria. The three affected relatives were heterozygous for the mutation. The alternative base was represented in bold.

Family G consisted of two affected sisters with two heterozygous variants in *NPHS2* (p.R138Q and splicing variant) in compound heterozygosity. Both sisters were diagnosed with primary SRNS and FSGS and developed multidrug resistance with an early age of onset (Figure 31).

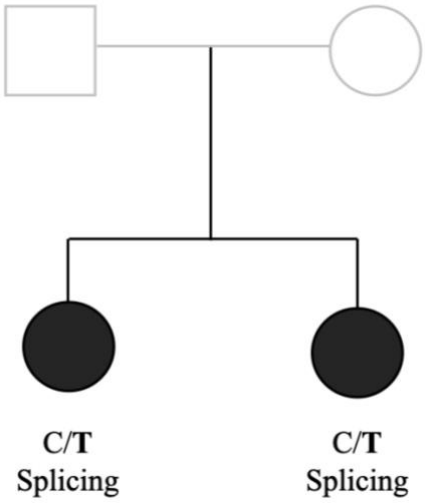


Figure 31. Segregation of the variants in *NPHS2* (p.R138Q and splicing variant) in Family G. Squares and circles indicate males and females respectively. The alternative base was represented in bold. Both siblings had the two variants in compound heterozygosity. DNA was not available for sequencing on the unaffected parents.

4.2.3 Previously described or novel variants in established genes that are inconsistent with previously reported mode of inheritance

A total of 22 candidate disease-causing variants (novel or previously described) in known SRNS genes with an inconsistent model of inheritance were detected in 20 primary SRNS cases. All cases had one variant except for S1656 that had two variants, one in *COL4A*1 and another in *NPHS1* and S0760 that had two heterozygous variants in NPHS1. All variants described in this section were found in the following genes: *NPHS1*, *NPHS2* and *COL4A1* (Table 8). The inherited conditions associated with mutations in *NPHS1* and *NPHS2* in relationship with nephrotic syndrome follow an autosomal recessive model (72, 73) and the variants described in this section were heterozygous. Therefore, these variants could be potentially contributing to the phenotype but the evidence to support causality is weaker than in the two previous sections, as the genotypes were not consistent with the known mode of inheritance of the genes. Some studies in *NPHS1* and *NPHS2* have suggested that patients with

heterozygous mutations have proteinuria but may respond to therapy and have a good long-term outcome (175, 176). However, further investigations are required to confirm the link between heterozygous mutations in *NPHS1* and *NPHS2* and the development of proteinuria. Interestingly, S0760 could have compound heterozygosity in *NPHS1* but one of the variants was synonymous (p.C528C) and would be considered neutral as it is not expected to have a damaging effect on the encoded protein. Additionally, *COL4A1* has been associated with nephropathy in an autosomal dominant pattern (177) although the clinical phenotype of S1656 did not match with the ones described for mutations in the gene (178). Consequently, these 20 cases remain unsolved.

4.3 Discussion

Whole exome and whole genome sequencing enabled the efficient screening of 67 genes that have been previously associated with nephrotic syndrome in a cohort of 422 individuals from different ethnicity groups. A total of 48 patients had one or two pathogenic or likely pathogenic variants that explained their phenotype. From those 48 patients, four probands have a pathogenic variant that explained their family phenotype in the following genes: *NUP107, LCAT, TRPC6* and *NPHS2*. Collectively, mutations in the established nephrotic syndrome genes collectively explained ~11% of the cases in the cohort. In contrast, previous genetic analyses of other SRNS cohorts used whole exome and also targeted sequencing strategies but had a smaller number of genes screened such as the one carried out by Sadowski CE et al (82). Although the number of genes screened was smaller (27 genes) they found a single-gene cause in 29.5% of their patients. This study was the first one to analyse a large international cohort of paediatric SRNS patients with 1783 individuals with familial disease (82). Since the study of Sadowski CE et al and most others have analysed exclusively

paediatric cohorts with familial disease, these results could reflect the difference in study population in comparison with the cohort studied in this dissertation. In view of the vastly higher incidence of known SRNS gene mutations in early life and familial cases, this would in part explain with the observed difference with the study cohort here since this was not exclusively paediatric, and also preferentially recruited sporadic rather than familial SRNS. Majority of samples with a previously established pathogenic or likely damaging variant in this cohort had a childhood onset, supporting previous findings of a higher incidence of monogenic disease in childhood (Figure 24).

Other recent studies such as the one carried out by M Wang et al used whole exome sequencing to screen for 165 genes associated with kidney disease and found rare variants in known disease-associate genes in 33.6% of their cases (179). However, the cohort studied was not exclusively comprise of primary SRNS patients but a histology compatible with FSGS (179). Therefore, secondary causes of FSGS such tubular disease were not rejected. As such, genes that would not normally be associated with primary SRNS were also significant in the burden test (179). Thus, differences in the clinical eligibility criteria could lead to the differences in the percentage of the cases explained by monogenic disease in other studies compared with the cohort examined here. Additionally, in this study cases that underwent whole genome sequencing as part of the BRIDGE study (*Rare Disease Pilot for the 100,000 Genome Project),* were previously screened for some established SRNS genes for this reason the recruitment was biased towards patients that did not have an obvious monogenic SRNS form. Overall, this cohort better represented the heterogeneity of SRNS in renal clinic populations (paediatric and adult) in clinics across the UK, as most previous studies selectively recruit patients with an especially high likelihood of a monogenic disorder.

Mutations were only detected in 13 of the 67 genes currently associated with nephrotic syndrome (Figure 25.A). *NPHS1* and *NPHS2* had the highest frequency of mutations, as has been described in other cohorts (60, 82).These findings highlight the heterogeneity of the disease and suggests that some of the variants or genes previously reported in the literature might be questionable in terms of nephrotic syndrome causality. For this reason, it is crucial to correctly discriminate between causal, uncertain significance or benign when the variant pathogenicity is assessed. During the course of this work, a number of different filters were carried out in order to generate robust findings and ultimately identify genuine causal variants. For instance, synonymous variants were considered as neutral because they are not expected to have a damaging effect on protein, although in some studies they have been shown to affect mRNA stability and protein expression (180). Prioritisation of variants by allele frequency was used to enrich potential pathogenic variants over likely neutral variants, due to the expectation that deleterious variants would be selected against (118). CADD score was also considered as a well-established guide to predict how damaging a variant might be. Overall, by removing a substantial number of variants from the analysis using the described cut-offs, could lead to failure of identifying a truly causal variant (false negative result) (119). Having analysed the causative variants identified in this report, as shown in Figure 25, it might be useful to identify future variants in novel genes by improving the design of cut-offs used to filter variants.

Apart from monogenic forms of SRNS, the risk alleles G1 and G2 in *APOL1* were also assessed. 72% of the patients with African descent had a risk allele in *APOL1* suggesting the high prevalence of G1 and G2 in individuals of African ethnicity with SRNS. These risk alleles are associated with moderated to severe FSGS and CKD (84, 87, 172). Therefore, the phenotype of these patients was very heterogeneous as

expected, with early and late age of onset and variety of long-term outcomes (Figure 25.B).

Despite stringent quality control steps and adequate filtering, the different coverage across established SRNS genes could have resulted in failure to identify causal variants in the cohort. As described in the methodology section in Chapter 2, in order to maximise the number of patients with SRNS, two datasets from different projects were combined and processed using the same pipeline. Thus, cases from the BRIDGE project (WGS) were sequenced at lower depth than the samples sequenced from KCL genome core (WES) (section 2.5). When studying monogenic causes of SRNS, coding variants were analysed rejecting noncoding variation as not all patients were whole genome sequenced and there are many challenges regarding the functional interpretation of noncoding variants (181). Furthermore, the variable exome capture technology led to differences in the sensitivity and specificity of the data. Other studies had reported variability in the coverage depth across the exome (182, 183). Thus, it is recognised that high-throughput sequencing technologies have limitations that could potentially lead to an impact in the variant identification.

Finally, this chapter illustrates the importance of genotype – phenotype correlation in ascertaining the true pathological significance of any mutation.

Chapter 5 – Rare genetic variants in novel candidate genes

5.1 Introduction

As previously described in chapter 4, the SRNS cohort was studied to identify the incidence of mutations in established SRNS genes as well as their patterns of inheritance. Single gene mutations appeared to again explain only a minority of cases (11%) in comparison with previously published literature (30%). As such, the available evidence pointed to more complex genetic architecture in the majority of patients. The 67 genes that are known to be mutated in SRNS were screened across the cohort in order to find previously described or novel pathogenic variants. Aside from monogenic inheritance detected in some familial as well as sporadic disease, potential risk alleles in predisposition genes such as APOL1 were also identified.

In this chapter novel and rare genetic causes of large effect in genes that have not been previously associated with SRNS were explored. Therefore, the aim was to identify the contribution (if any) of novel genes to SRNS in both familial and sporadic cases that had not been solved through detecting causal mutations in established SRNS genes.

Cases where no pathogenic or likely pathogenic variants had been detected in established SRNS genes were selected for further study. For each, the genetic variation profile was built based on exome sequencing data and then filtered searching for likely damaging variants in a wider spectrum of genes chosen either through their association with glomerular kidney disease and/or genes significantly enriched in podocytes. Different methods of family analyses including parametric linkage analysis or

nonparametric linkage analysis were performed to identify genomic regions potentially harbouring causal variants. Finally, a case-control study was undertaken in SRNS samples of European ethnicity to identify novel disease risk and/or causative variants of potential therapeutic relevance.

5.2 Identifying likely damaging variants per sample in novel genes

Unsolved cases within the SRNS cohort underwent data processing using the pipeline described in the section 2.5. A variant profile was generated for each sample with annotated genetic variants that passed the quality control filters. This contained variants detected within coding regions and intronic variants in regions in the immediate vicinity of exons likely to participate in splicing. This generated individual variation profiles containing approximately ~16,000 variants each annotated using standard characteristics of genetic variation such as allele frequency, protein consequence and pathogenicity scores (outlined in section 2.5.4).

To identify deleterious genetic mutation in novel genes, variants were filtered using the cut-offs described in Chapter 4 based on characteristics of likely causal variants. As previously mentioned, solved cases with a causal variant or variants identified were excluded from this analysis, leaving a total number of 374 residual cases explored to analyse. This prioritised rare variation (AF<0.001), variants with a CADD score higher than 15 and rejection of synonymous variants. Only variants located in ±3 bp from the exons were selected. Heterozygous variants that were found in more than 9 controls from the in-house dataset (analysed using the same pipeline) or as homozygotes in more than 5 controls were also rejected, especially if this was

corroborated in control databases such as gnomAD. Variants of potential interest were extracted using two exploratory gene lists: one was a list of 3,019 genes expressed in podocytes generated using the raw data from the study performed by H Yu et al (DNA microarrays of mouse podocytes and extracting the human orthologues) (184) the other a list of 316 genes associated with the HPO term 'Nephrotic syndrome' in OMIM (https://hpo.jax.org/app/). After filtering, there were approximately ~380 variants of interest per sample. A further step of filtering for variants within the same gene, across samples was also performed. The gene with the most variation across samples was *AHNAK2*, a gene mostly associated with neuromuscular disorders and cancer. Although there are no direct links between *AHNAK2* and SRNS to date, it is expressed in podocytes and paradoxically lies within the critical region of chromosome 14 harbouring the gene *INF2* which is associated with autosomal FSGS (185), with both genes located within 400kb of each other. A very tenuous phenotypic link is that both *INF2* and *AHNAK2* are also associated with a rare neuromuscular disorder, Charcot Marie Tooth which in turn can be associated with SRNS (186, 187). Recent studies have also found that *AHNAK2* is a prognostic marker for clear cell renal carcinoma (188).

As such, although this analysis was informative, the evidence that any of these genes including *AHNAK2* may be associated with SRNS was not robust and therefore inconclusive. Accordingly, other methodologies such as family analysis and case control study were performed to generate more evidence in support of novel genetic causes of SRNS.

5.3 Family analysis

Positional cloning combined with linkage and homozygosity mapping were the first approaches to be used in familial SRNS to successfully identify genes responsible for the disease (74, 189). These analyses focused on the search of genetic markers across the genome of affected relatives in an attempt to narrow down the causal variant responsible for the phenotype of the family. Segregation and linkage studies have been particularly informative in finding the genomic regions of highly penetrant disease-causing variants as they should be presented in the affected relatives but not in the unaffected. However, there are challenges when using this approach as there is no guarantee that the family of interest has a fully penetrant mutation and that there are enough individuals to allow a meaningful statistical analysis. Thus, when families are smaller and/or not sufficiently interbred to allow homozygosity mapping, or the causal mutation is not fully penetrant such that affected members may not be initially obvious and included incorrectly in analysis as unaffected individuals, linkage studies may be unsuccessful until this point is reached.

In the previous chapter, only four families (Family 5, B, C and G) out of the total 14 families sequenced in the study cohort were solved by detection of mutations in established SRNS genes. Therefore, this chapter focused on the remaining 10 families that did not have causal variants in the coding regions of the established SRNS genes (Family 1, 2, 3, 4, 6, 7, A, D, E and F). Three main strategies were applied to search for potential novel candidate genes. First, all families were studied by segregation analysis extracting a list of likely damaging variants shared by affected relatives of the same family. Additionally, the results for each family were compared to see if there were any recurrent variants in the same genes among families with comparable

phenotypes. Parametric linkage analysis was used to identify regions of the genome that might harbour a causal variant that could explained the SRNS dominant trait in Family A. Finally, out of the 10 families, 7 met the criteria (at least two affected siblings) to be evaluated by nonparametric linkage analysis to explore allele sharing across the affected individuals of the families.

5.3.1 Segregation analysis

Segregation analysis explores if the pattern of a phenotype within a family is consistent with genetic inheritance of disease. All 10 unsolved families with no causal mutation detected were studied by segregation analysis to identify whether cosegregation was present in the affected members of each family. To search for potential disease-linked variants, those shared between affected relatives yet absent in the unaffected were selected, filtered by allele frequency and scored for significant protein alteration, as described in the section 2.9.1.1. Each family was explored under both a dominant and recessive model of inheritance (Table 11).

As shown in Table 11, most genes presented variation in just one family, however a minority did exhibit recurring variation shared across the families, namely *MUC4* which showed multiple variants in six families, followed by *AHNAK* that had variants found in five families, and *AHNAK2* with variants found in 4 families (Table 12). However, an identical variant was detected in *AHKAK* in multiple families. This was likely artefactual as it is not expected that the same variant would be recurrent as the families were not related. Other genes only showed variation across three families such as *PRDM9, MYO15A* and *FLG2*. Finally, these genes were explored in context of kidney expression and kidney disease. As discussed previously, *AHNAK2* is expressed in podocytes according to the list generated by H Yu et al (184) and has

been associated with renal carcinoma (188). *MUC4, FLG2, MYO15A* and *PRDM9* were highly expressed in kidney but were not associated with SRNS; *MUC4* is an oncogene, *FLG2* is primarily associated with skin disorders, MYO15A is a deafness gene including Usher syndrome (although this is associated with kidney disease this is ciliopathy rather than SRNS) and *PRDM9* is again associated with a number of rare syndromes, such as velocardiofacial syndrome, but none of these had SRNS as a feature. As such none of these could be placed as plausible candidates for familial SRNS.

Table 11. Number of variants segregating in each family. Variants segregating in the affected relatives of each family were extracted under recessive and dominant model of inheritacne.

Family ID	Variants segregating (dominant)	Variants segregating (recessive)	Potential gene candidates
1	38	1	*MUC4, AHNAK2*
2	34	6	*MUC4*
3	46	3	*MUC4, PRDM9*
4	34	2	*AHNAK2*
6	492	15	*MUC4, AHNAK2, MYO15A*
7	107	4	*AHNAK2, MYO15A, PRDM9*
A	-	-	-
D	127	6	*MUC4*
E	141	1	*MUC4*
F	398	15	*MYO15A, PRDM9*

Table 12. Genes with the highest variation recurrence across families.

Gene	Number of families with variation recurrence
MUC4	6
AHNAK	5
AHNAK2	4
PRG4	4
FLG2	3
MYO15A	3
PRDM9	3

Interestingly, Family A did not exhibit any coding variants that were segregating within the affected members of the family Therefore, this multigenerational family with a clear pattern of autosomal dominant inheritance, was selected for more detailed analysis (Table 11).

5.3.2 Parametric linkage analysis

Analysis of families with individuals from different generations that have accurate genotype and phenotype information, can be used to establish if variants segregate with disease using the suspected model of inheritance via linkage analysis. In Family A, no potential pathological variants were detected by segregation analysis that met those criteria in the coding regions (Table 11). Notwithstanding the limitations in detecting rare genetic variants in areas of the genome by WES analysis (190). This result potentially supports a hypothesis that the variant causing SRNS in this family might be located within a non-coding region.

Detailed family history and phenotypic examination of Family A revealed four generations of affected individuals with DNA available in three generations for sequencing, and a pattern of inheritance compatible with an autosomal dominant model (Figure 21). The age of onset of symptomatic SRNS as measured by development of nephrotic range proteinuria and end stage kidney failure in affected individuals ranged from early teens to mid-thirties; however significant abnormal urine protein loss indicative of impending disease could be detected in affected individuals from late childhood, either through a urine test or historical examination of medical records. Unaffected individuals had undetectable protein loss (as expected for a normal individual) as measured by protein and albumin creatinine ratios in the urine even in middle age. All affected individuals presented FSGS, with histological

confirmation in five individuals from three generations by kidney biopsy. 12 individuals from the family including the proband underwent exome sequencing (seven affected and five unaffected). Initially, sequences were analysed to check for the presence of likely pathogenic variants in the coding regions that were present in the seven affected individuals but not the five unaffected, but no variants that met the criteria were detected.

In view of this result, parametric linkage analysis was then performed using MERLIN software under the assumption of autosomal dominant inheritance. MERLIN uses a fast algorithm based on sparse trees to represent the flow of genes through pedigrees. The evidence for linkage is assessed by computation of a heterogeneity LOD score. A positive LOD score indicates excess allele sharing among affected individuals and a negative indicates less than expected allele sharing. For this family, a region in chromosome 2 located between co-ordinates 180,810,180 to 180,835,792 with a LOD score greater than two (with a maximum LOD score of 2.1154) was identified (Figure 32) (Table 13). Thus, haplotype analysis of the segregating genomic region in the pedigree indicates that there is not consistent cosegregation with an extended haplotype, though we cannot exclude the possibility of a short shared haplotype. Given the clinical presentation of the disease it appears more likely that there are phenocopies or non-penetrance. The linkage peak in chromosome 2 contained one gene *CWC22*, that has not been previously associated with kidney disease although is highly expressed in the nervous system and kidney. *CWC22* associated with rare autosomal cancer and neurological syndromes (191, 192).

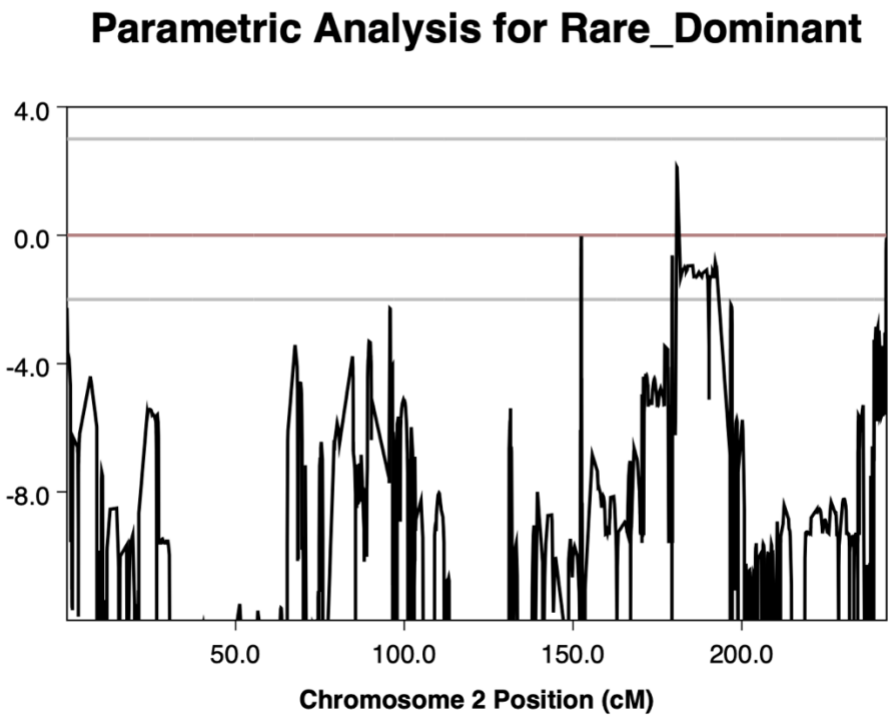


Figure 32. Parametric linkage analysis of Family A. The parametric model used was for a rare dominant trait. Plot shows multipoint LOD scores from chromosome 2. The maximum LOD score was 2.11. Data and plot generated by MERLIN.

Table 13. Variants with the highest evidence of linkage in the region in chromosome 2.

CHR	POS	LABEL	MODEL	LOD	ALPHA	HLOD
2	1.8081	chr2 180810180 C T	Rare Dominant	2.1154	1	2.1154
2	1.8081	chr2 180810264 A T	Rare Dominant	2.1154	1	2.1154
2	1.8081	chr2 180810358 C T	Rare Dominant	2.1154	1	2.1154
2	1.8081	chr2 180810443 C T	Rare Dominant	2.1154	1	2.1154
2	1.8084	chr2 180835589 A G	Rare Dominant	2.1069	1	2.1069
2	1.8084	chr2 180835792 T A	Rare Dominant	2.1069	1	2.1069

5.3.3 Nonparametric linkage analysis

From the remaining ten families where no causal mutation was identified in the coding regions of established nephrotic syndrome genes, seven (Family 1, Family 2, Family 4, Family A, Family D, Family E and Family F) were also analysed by nonparametric linkage analysis using MERLIN as described in the section 2.9.1.3. This methodology does not require any information about the model of inheritance and uses genotypes from affected siblings from multiple families to explore markers that might be shared among families more often than would be predicted randomly. Thus, it requires at least two affected siblings per family to be informative and only seven families out of the ten met those criteria.

Filtering the genetic data by genotype quality and allele frequency as described in the methodology (section 2.9.1.3), from a total of 100,717 possible sites shared across the seven families just only 27,058 genotypes were kept. Furthermore, because the standard linear model used by MERLIN is too conservative, the analysis was also repeated using an exponential model to accommodate the small number of families and large increases in allele sharing expected due to the nature of the disease and potential for high penetrance mutation. The pedigree structure was 41 individuals (20 females and 21 males) with a SRNS prevalence of 52.5%. The average of the family size was 5.86 and the generation average was 2.14. The maximum possible LOD score using the exponential model was 4.0 and the minimum was -2.473. None of the markers met the criteria for genome-wide significance (a LOD score of 4.0). However, some regions in chromosome 2 reached a LOD score of 2.811 (from position 186,625,770 to 192,701,393) and chromosome 7 reached 2.319 (from position 100,551,692 to 100,553,073) which could be considered suggestive evidence of

linkage (Figure 33). The linkage peak in chromosome 2 was broad and contained 40 genes. To prioritise the search in chromosome 2, genes that are expressed in podocytes were selected resulting in 7 genes from the total 40 (Table 14).

The peak in chromosome 7 was smaller and contained only two genes *MUC3B* and *MUC3A*. The mucin genes are part of a family of epithelial glycoproteins and transmembrane ones are *MUC1, MUC3A* and *MUC3B*. *MUC1* has been associated with kidney disease and FSGS (193), whereas *MUC3A* has been associated with renal carcinoma (194) and *MUC3B* has not previously been associated with kidney disease, though does have a role in kidney development (195).

Then, the two variants with the highest LOD score in both peaks (position 191,224,981 in chromosome 2 and 100,552,017 in chromosome 7) were investigated to identify which families were contributing the most to the linkage signals running MERLIN with the '--perFamily' option. The family that contributed the most in both regions was Family A. In chromosome 2, Family E and Family F where the ones with the highest LOD score after Family A and Family 1 had a negative LOD score (Table 15). In chromosome 7, all the families apart from Family A reached a LOD score of 0.3 except from Family 2 that had a negative score (Table 16).

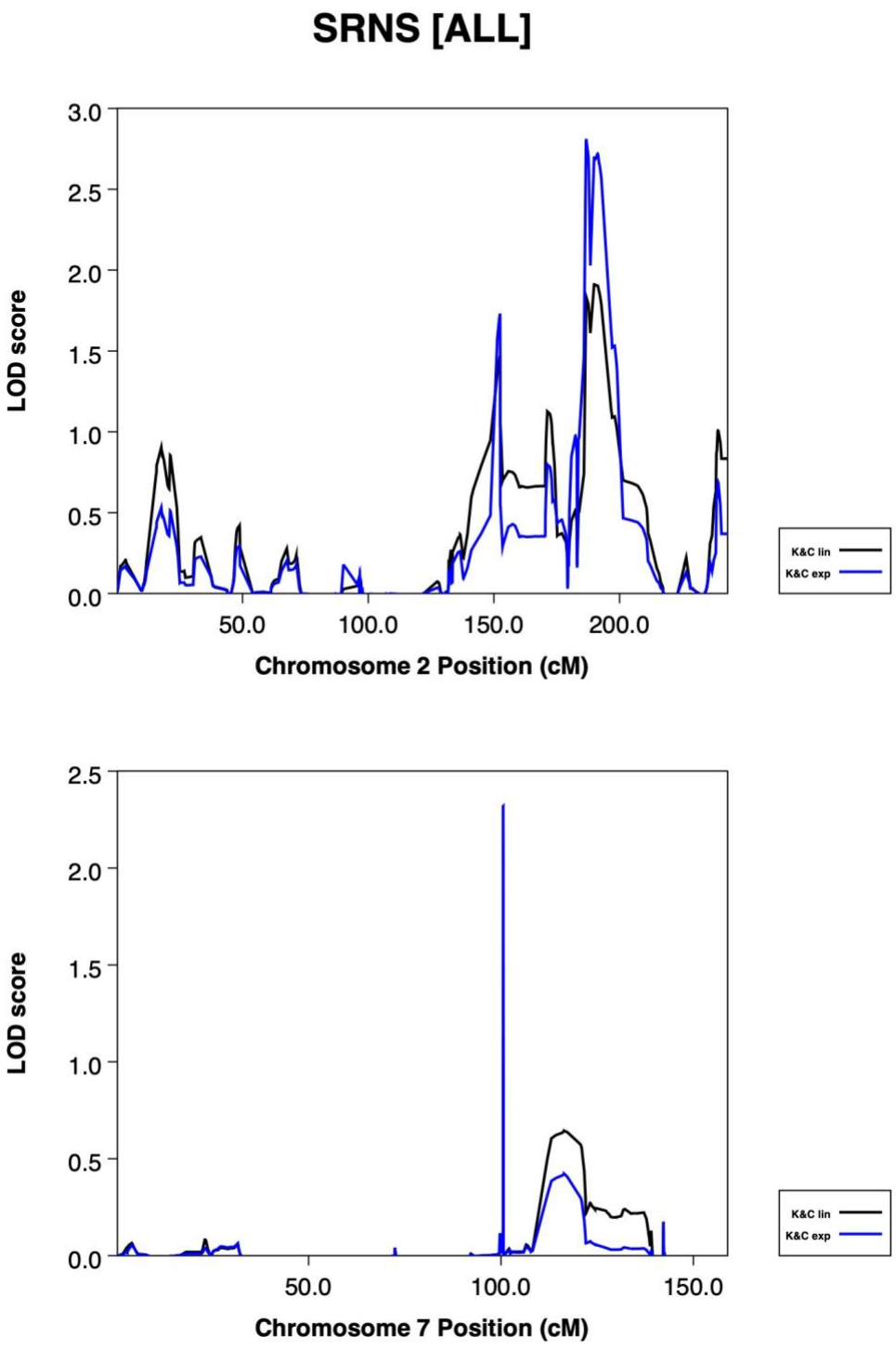


Figure 33. Nonparametric liankge analysis results from the chromosome 2 and chromosome 7. The linear (black line) and the exponential (blue line) model based on the work made by Kong and Cox are shown in the plots. Regions in chromosome 2 and chromosome 7 reached suggestive evidence of linkage.

Table 14. Genes from the linkage region in chromosome 2 that are expressed in podocytes.

CHR	STRAND	START	END	GENE
chr2	+	189856394	189859058	*COL3A1*
chr2	+	191273080	191367041	*MFSD6*
chr2	+	191523883	191557492	*NAB1*
chr2	+	191745546	191761227	*GLS*
chr2	+	192110106	192290115	*MYO1B*
chr2	-	188206689	188313021	*CALCRL*
chr2	-	188328957	188419219	*TFPI*

Table 15. Nonparametric linkage results per family in chromosome 2.

FAMILY	LOCATION	Z-SCORE	pLOD	DELTA	LOD
1	chr2 191224981	-0.000009	-0.000003	-0.707107	-0.000003
2	chr2 191224981	1.305773	0.284051	0.707107	0.284052
5	chr2 191224981	0.989302	0.230332	1	0.298701
A	chr2 191224981	4.509114	0.62205	1.276731	0.829749
D	chr2 191224981	0.710363	0.176757	1.224745	0.271845
E	chr2 191224981	1.340723	0.289596	0.745356	0.300881
F	chr2 191224981	1.4142	0.301028	0.707107	0.301028

Table 16. Nonparametric linkage results per family in chromosome 7.

FAMILY	LOCATION	Z-SCORE	pLOD	DELTA	LOD
1	chr7 100552017	1.414209	0.236417	0.707107	0.301029
2	chr7 100552017	-1.412788	-0.557203	-0.707107	-0.300811
5	chr7 100552017	0.999988	0.179438	1	0.301027
A	chr7 100552017	4.943077	0.547643	1.276731	0.863975
D	chr7 100552017	0.861127	0.158532	1.224745	0.31274
E	chr7 100552017	1.341402	0.226929	0.745356	0.300991
F	chr7 100552017	1.414213	0.236418	0.707107	0.30103

5.4 Case-control analysis: gene-based burden test

Association analysis is a very useful strategy to identify novel disease variants in the absence of groups of individuals from the same pedigree. Case-control studies require the selection of cases with the same phenotype, as well as the collection of controls without the phenotype. A gene-based burden test analyses whether there is an

enrichment of rare variants (low frequency) in any one gene in cases versus controls. Despite the successful findings made by such studies, there are several difficulties associated with the collapsing of rare variants within a gene, as it raises the problem of how to deal with neutral variation that might occur in pathogenic genes in both cases and controls. Additionally, extensive quality control strategies must be considered to avoid systematic biases and false positives when studying sequencing data from different sources.

5.4.1 Study design considerations

Rare alleles responsible for Mendelian disorders have high penetrance and affect the genetic fitness of the carrier individuals. Thus, rare causal variants are strongly influenced by natural selection and are less likely to be transmitted to subsequent generations. In order to run successful single-variant association tests for rare diseases the sample size has to be sufficiently large and the effect of the causal variant has to be strong. Because of the rare disease status of SRNS and as a consequence the limited sample size of the cohort a multi-variant collapsing method was chosen instead (114).

Gene-based burden testing combines information of genetic variation across cases and controls within a gene. This strategy overcomes the limitations of single-variant association tests but relies on most variants included in the analysis to be causal. While there is no definitive procedure to distinguish between harmful and non-harmful genetic variation, certain filters can be introduced to enrich for variants with a similar profile to pathogenic events. Therefore, the burden test relied on several assumptions made about the characteristics of the variants studied such as frequency, functional prediction and protein consequence. This study assumed that causal variants were located in coding regions of the genome, rejecting regulatory elements and other non-

coding variants outside of the exome. Synonymous variants do not usually have a damaging effect on the protein function, aside from rare cases of functional synonymous changes introducing splicing. Due to the negative selection pathogenic variants were assumed to be rare in the control population and common variants were rejected. Variants were filtered by allele frequency depending on the model of inheritance (dominant or recessive). In addition, the CADD score was used to predict the damaging effect of a variant in the protein function. The unit used for the analysis was a gene, assuming all sections of the gene are equally important, when it is recognised that certain exons that encode protein domains are directly responsible for the protein function and may therefore cluster pathogenic mutations (196) (Figure 34).

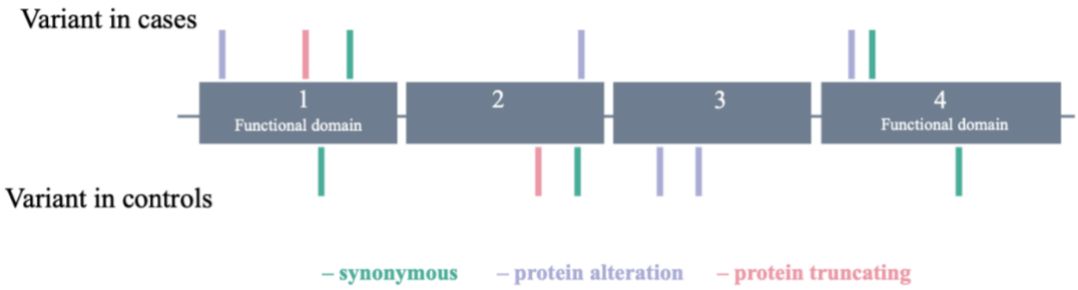


Figure 34. Hypothetical pattern of variation in cases and controls across a gene. The diagram shows a gene with four exons where two of them contain a functional domain important for the protein function. This gene harbor pathogenic variants in the function domains in the cases but not in the controls although they both have same absolute number of variants.This example higlights the difficulties and problems faced when performing a gene collpasing burden test.

Quality control steps were also performed to detect sequencing errors and variation that was poorly captured in one or more of the datasets simply through differences in sequencing depth. Specifically combing exome data with lower coverage whole genome data requires a joint analysis to identify systematic differences. Additionally, some extra filtering recommendations were followed to remove batch effects caused by different experimental conditions including different exome capture kits used over the time patient recruitment occurred (explained in section 2.4.2 and 5.4.3).

5.4.2 Joint calling across samples

To minimize technical artefacts, the case-control analysis was run using variants that were called across all samples by performing a variant calling jointly. This strategy improved the sensitivity and specificity of the variant calling process allowing early detection of errors or false positives. Joint calling distinguished whether a variant call (reference or alternative) was made at that position for all samples. Thus, it detected variants that were not seen in other samples because no call was made at the variant location for technical reasons and only variants that were called in at least 90% of samples were used in the gene-based burden test. Moreover, an advanced statistical approach called variant quality score recalibration (VQRS), was used to identify the technical profile of variants using machine learning algorithms and remove systematic sequencing errors (explained in section 2.5.3.1).

5.4.3 Data filtering

Before embarking on downstream genetic analysis, quality control procedures were performed in the multi-sample VCF to ensure high quality sequencing data. Thus, the coverage, missingness and the ratio of heterozygous and homozygous variants in each sample was investigated to detect outliers. Additionally, samples were selected depending on their ethnicity and relatedness (with other cases) to include the maximum number of unrelated cases in the analysis.

Ancestry matching was performed to avoid detecting rare variants that might appear pathogenic in cases through enrichment in a particular ethnicity rather than any real association with the disease. Therefore, only cases and controls that cluster via principal component analysis with individuals of European ancestry were selected. In the absence of appropriate cohorts of ethnically matched controls beyond those

reflecting a European ancestry, analysis was carried out just in Europeans as this was also the largest ethnic group (Figure 35.A). Relatedness analysis was conducted generating a kinship matrix using methodology described in 2.7.1 section. The kinship coefficient is the pairwise relationship between two samples and is achieved by adding the scores of any two samples together. Thus, identical samples and identical twins will score 0.5, sibling score approximately 0.25 and cousins score 0.125. For closely related family members (parents, sibling, cousin, etc.) only one sample was included from the entire family (proband). From all pairs with a kinship coefficient higher than 0.12, one sample of the pair was removed (Figure 35.B). The genotype missingness rate was calculated for each sample using BCFTools (197), to show the proportion of variants for which no genotype was called (Figure 35.C). This led to 8 controls and 2 cases being rejected from the analysis because of their genotype missingness rate. Possible sample contamination was suspected in 11 controls and 34 cases due to high or low levels of heterozygosity (Figure 35. D), and these individuals were also excluded from further analysis.

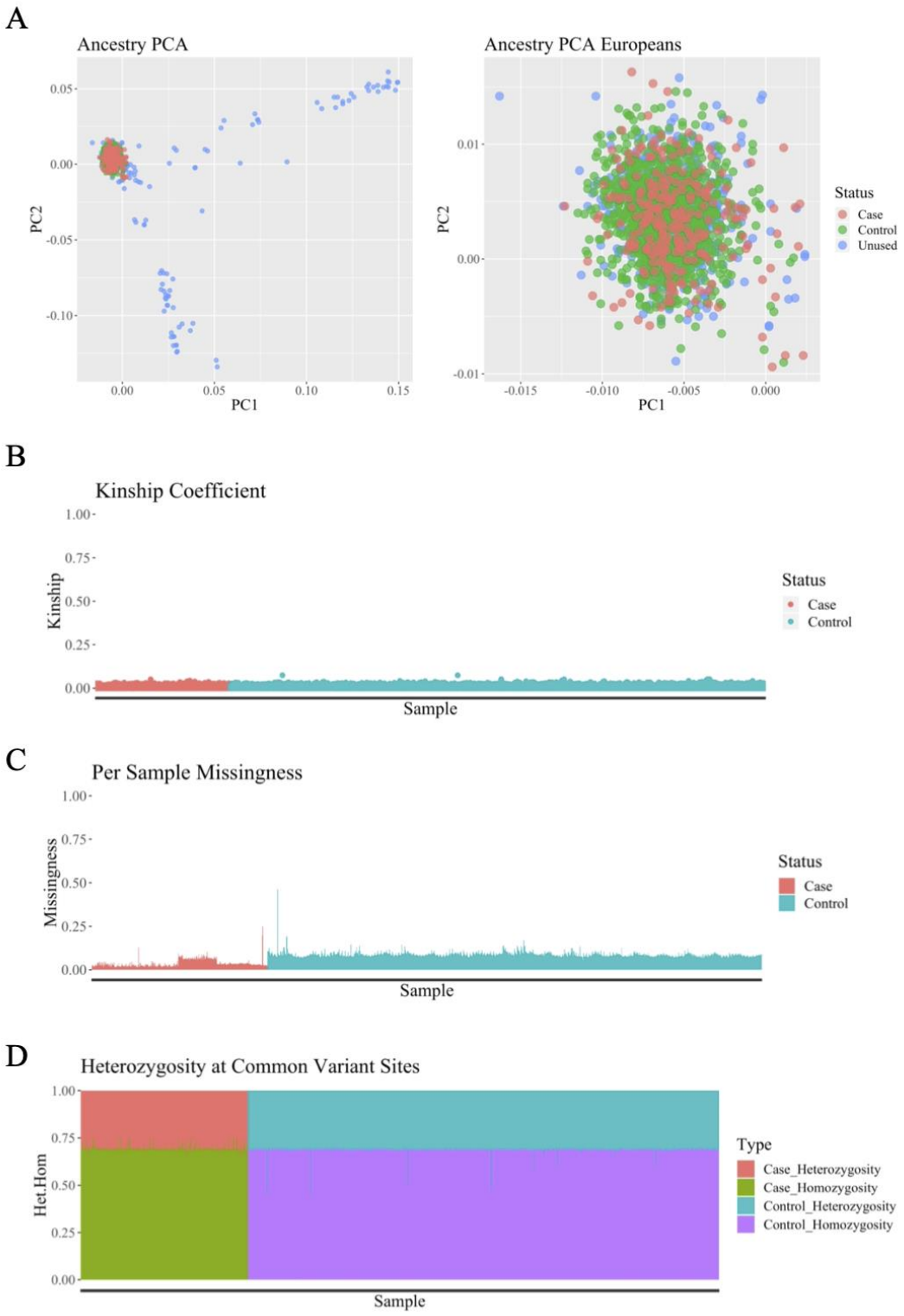


Figure 35. Case-control analysis quality control steps. (A) European cases and controls were selected via principal component analysis by setting up the cut-off in PC1<0.0025. Cases and represented in red and controls in green. (B) Kinship coefficients for all samples showing relatedness within the cohort. (C) Missingness rates of all samples from 0 to 1.0. (D) Heterozygosity/homozygosity ratio at common variants within the cohort.

The proportion of reads for each sample was calculated providing information on the exact depth of coverage. As expected, when combining WES and WGS datasets, the coverage of cases differed depending on the type of sequencing (Figure 36). WES cases and all controls had higher coverage than the WGS cases. Due to the increased sensitivity to detect variation across datasets with higher coverage and decreased specificity to evade sequencing errors in lower coverage datasets, an additional filter was applied to reject genotype missingness at a group level. Thus, sites with a missingness rate higher than 10% in the WES cases, WGS cases and controls were filtered separately by group as shown in Figure 37. Then the intersection of variants across all groups was selected and filtered by a missingness rate higher than 10%. Variants in low complexity regions were rejected. Biallelic variants were selected and filtered by depth (DP) and genotype quality (GQ). Furthermore, all variants were annotated with allele frequencies from gnomAD and CADD score and common variants were filtered out (AF<0.01). Correspondingly, from a total of possible sites 201,803 shared across all samples just only 169,628 genotypes were kept for further analysis.

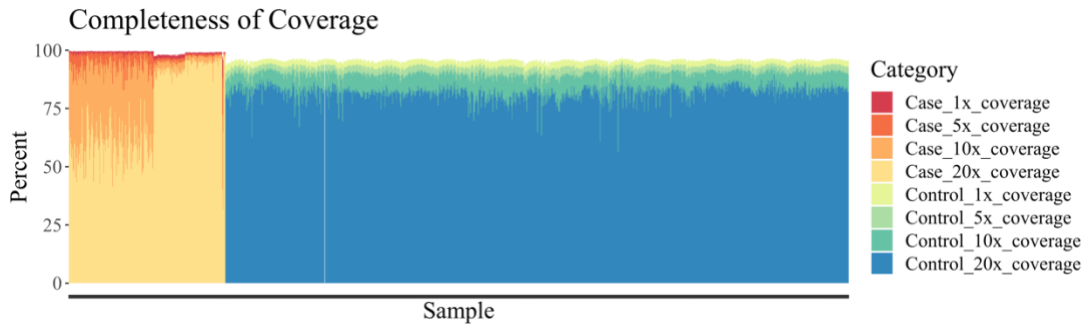


Figure 36. Coverage distribution across cases and controls. Each bar represents and individual samples and the percetange of genotypes with at 1X,5X, 10X and 20X depth coverage.

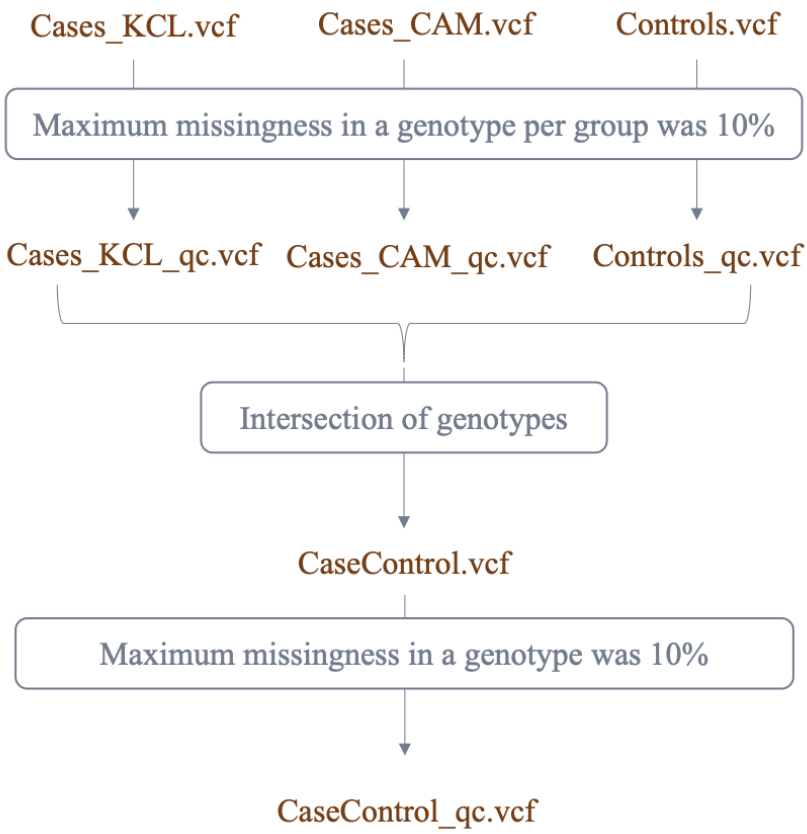


Figure 37. Filtering pipeline to extract genotypes with no missingness. The three original datasets were filtered by missingness and then the intersection of genotypes across all samples underwent the same filter too.

Finally, samples with known causal variants or variants that were highly likely to be causal based on previously published work were excluded. In particular, some samples had novel or non-published variants that were similar to previously confirmed variants in SRNS associated genes. Often, these novel variants were in a known hotspot for pathogenic mutations within the gene or were predicted to have an deleterious effect on the function of the gene (e.g. truncating a protein before a crucial domain). Consequently, 48 cases from the test cohort with pathogenic variants identified in the section 4.2 were excluded from the case-control analysis.

5.4.4 Results

Following the quality control filters to ensure that samples were appropriately selected for further analysis, there were a total of 245 cases and 970 controls that passed all the cut-offs as explained in the previous section. Then, a gene-based rare variants burden test was carried out by EPACTS framework as explained in the materials and methodology section (2.9.2.1). Collapsing all variants across a gene together, allows testing for association under the assumption that all variants have a consistent direction of effect. For this analysis, two burden tests were performed using different methodologies to select rare variants. In the first test, all rare variants observed in one gene in cases and controls were aggregated and classified by three categories (alteration, truncation and synonymous) and model of inheritance (dominant and recessive) (Figure 38.A). In the second test, all rare variants observed in one gene in cases and controls were aggregated and filtered by CADD score (>10) and model of inheritance (dominant and recessive) (Figure 38.B). Significance of the burden association was determined with one-tailed Fisher exact test (section 2.10.1), which test for an excess of variants in the case group versus the control group, under the assumption that rare functional variants in Mendelian diseases are damaging and not protective. Every gene containing one or more potential causal variants was tested for association. Genes were prioritised for downstream investigation based upon the evidence of association (p-value).

In the first test, there were three genes that reached burden significance ($p<2.5x10^{-6}$) (section 2.10.6). *FLG2*, the gene with the highest association, was significant under a recessive model of inheritance when the synonymous and alteration variants were aggregated and compared between cases and controls. *FLG2* protein expression is

more prominent in the tubular epithelium rather than glomeruli, however has been described as a glomerular extracellular matrix protein and recently ascribed a putative role in glomerular injury through participation in pathways leading to damage and FSGS (198). The synonymous variant group was performed to form a negative control as they are not expected to be pathogenic (except rare exceptions) and the number of synonymous variants should be similar in cases and controls. Here, the same synonymous variant 1-152325465-G-A was found in 5 cases and the 1-152324547-C-T in 4 cases suggesting the signal in FLG2 could be artefactual as it would be unusual for the same rare variant to occur across unrelated cases (Figure 38.A) (Table 17). Additionally, they did not pass the filter criteria for the genotype quality metrics in gnomAD (section 2.5.4.1). The other two genes that appeared significant within the recessive model were *ZNF257* and *ZNF780B*, belong to a family of Kruppel-like zinc finger proteins. While both are expressed in kidney, this was more in tubular epithelium than podocytes or glomeruli. There was no known association between kidney disease and/or SRNS with *ZNF257*, which appears to function more as an oncogene, or *ZNF780B*. Similarly to FLG2, signals of *ZNF257* and *ZNF780B* were driven by two variants in compound heterozygosity present in 8 cases. The genes with a suggestive p-value ($p>2.5x10^{-6}$), the only one that was associated with a kidney phenotype was *CR1L* which encodes Complement Component Receptor 1-Like Protein (Table 17). *CRIL* is expressed in glomeruli and also in podocytes and appears to participate in glomerular injury pathways as deficiency of Crry, the murine orthologue of *CR1L* (199). Nonetheless, the relevance to SRNS remains tenuous as complement is not normally involved in the pathogenesis of either FSGS or other SRNS related glomerular phenotypes.

When the gene burden test was repeated this time filtering by CADD score, only one gene, *ZNF780B*, reached the burden significance under recessive model in common with the previous test. However, *ZNF780B* has not been associated with any kidney phenotype and in addition, the discovery of two compound heterozygous variants shared across 8 unrelated cases (19-40541898-A-C and 19-40541912-G-C) suggested that the signal was not informative and it was unlikely that this was not a candidate (Figure 38.B) (Table 18). Furthermore, both variants were considered 'benign' according to the Polyphen-2 score although they had a CADD score higher than 10.

From the genes with a suggestive p-value in this test the only genes associated with a kidney phenotype were *APLNR* and *FOXL1* (Table 18). *APLNR* is not expressed in podocytes or kidney but encodes an adiopokine that when is overexpressed can injure podocyte, interfering with the autophagy (200, 201). *FOXL1* is highly expressed in podocytes and it has been associated with renal carcinoma (202, 203).

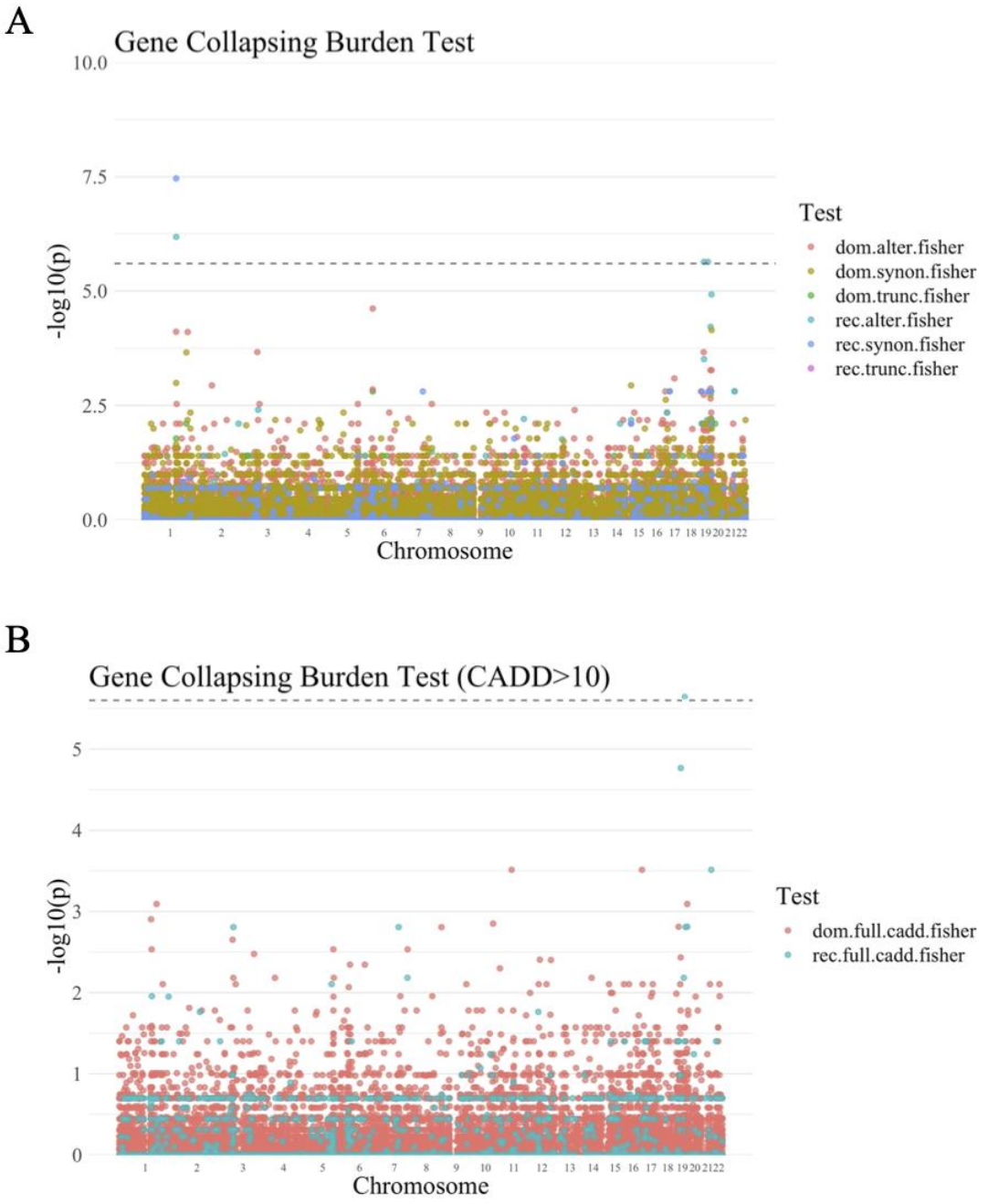


Figure 38. Summary of the results from the two gene-based burden tests by chromosome. Unrelated European cases (n=245) and controls (n=970) were compared by two burden tests. In both plots, each gene was tested for association by one-tailed Fisher's exact test. The x axis represents the position on each chromosome from the p terminus to the q terminus and the y axis the p-values on a logarithm scale. The dashed line represents the threshold for gene significance (p-value<2.5×10^{-6}). (A) The first burden test all rare variants across a gene were collapsed and classified by three categories: alteration, truncation and synonymous; and by model: recessive or dominant. (B) The second burden test all rare variants observed in gene were aggregated and filtered by CADD score (>10) and model of inheritance.

Table 17. Results from the burden test. Table is sorted by groups and p-value.

Test	Gene	p-value	Effect Size	Variant Cases	Variant Controls
Recessive Model					
rec.synon.fisher	*FLG2*	3.44E-08	50.362	12	1
rec.alter.fisher	*FLG2*	6.52E-07	8.4926	16	8
rec.alter.fisher	*ZNF257*	2.30E-06	inf	8	0
rec.alter.fisher	*ZNF780B*	2.30E-06	inf	8	0
rec.alter.fisher	*ZNF586*	1.18E-05	inf	7	0
rec.alter.fisher	*ZNF845*	6.03E-05	inf	6	0
Dominant Model					
dom.alter.fisher	*DPCR1*	2.41E-05	4.8018	17	15
dom.synon.fisher	*ZNF544*	7.09E-05	16.493	8	2
dom.alter.fisher	*FLG2*	7.73E-05	2.9487	26	38
dom.alter.fisher	*CR1L*	7.83E-05	28.773	7	1

Table 18. Results from the gene burden test filtering by CADD score. Table is sorted by groups and p-value.

Test	Gene	p-value	Effect Size	Variant Cases	Variant Controls
Recessive Model					
rec.full.cadd.fisher	*ZNF780B*	2.30E-06	inf	8	0
rec.full.cadd.fisher	*ZNF257*	1.71E-05	33.015	8	1
rec.full.cadd.fisher	*KRTAP10*	0.00030697	inf	5	0
Dominant Model					
dom.full.cadd.fisher	*APLNR*	0.00030697	inf	5	0
dom.full.cadd.fisher	*FOXL1*	0.00030697	inf	5	0
dom.full.cadd.fisher	*ANGPTL1*	0.00080872	9.5761	7	3
dom.full.cadd.fisher	*ZNF816*	0.00080872	9.5761	7	3

5.4.5 Gene-set analysis: hypergeometric test

The genes with a suggestive p-value in the burden test when filtering variants by pathogenicity were investigated to determine if they were previously associated with kidney disease (such as CKD) or highly expressed in the kidney (specifically podocytes). Therefore, the overlapping probability of the genes obtained from the burden test with the gene candidates acquired from other kidney studies was performed by a hypergeometric test. Since the list of genes expressed in podocytes that was previously used to extract potential novel variants in the section 5.2 was very broad (a total 3,019 human orthologs expressed in mouse podocytes) and list of 673 genes highly enriched in podocytes was chosen. This list was generated from a single cell RNA-seq experiment of human kidney tissue made by Gillies CE (64). Additionally, a list of 104 replicated loci that are relevant for kidney function in CKD in European subjects was also used. The resource was obtained from a GWAS meta-analysis of more than 1 million individuals with CKD by Wuttke et al (204).

There were 15,615 genes tested in the gene collapsing burden test filtered by CADD score (CADD>15) and 256 genes had a p-value <0.05. The overlap of the 256 genes was compared with the two previously described lists. The overlap with the list of podocytes enriched genes was 12 genes and with the list of CKD associated genes was 1. None of the hypergeometric test had statistical significance. (Table 19).

Table 19. Parameters used in the hypergeometric test. The P-values were generated using a one-sided test.

List	*K*	*k*	*n*	*N*	P-value
Podocyte enriched genes	256	12	673	15,135	0.14
CKD associated genes	256	1	104	15,135	0.18

5.5 Discussion

This chapter investigated potential novel genetic mechanisms contributing to SRNS by studying individuals for whom no causative mutation in the established genes have been identified. This comprised individuals with both familial and sporadic SRNS. Specifically, rare genetic variation that could be potentially pathogenic was explored using different methodologies including linkage-based analysis in familial SRNS and gene-based burden tests in sporadic cases. To narrow down the search for rare variants with a similar profile to causal mutations and reduce the chance of false positives, filters such as allele frequency and pathogenicity scores were implemented within analytical pipelines. Additionally, for the identification of likely damaging variation per sample, variants in genes associated with kidney disease and/or expressed in podocytes were selected.

The advantages of using family-based studies to identify new causal genes in a rare disease such as SRNS are the common genetic background shared by the relatives which enables to narrow down the search for potential candidates and the fact that families might undergo through similar environmental factors counteracting confounding effects. However, success of family analysis is heavily dependent on the recruitment process and factors such as pedigree size. It is critical to maximise the participation within the family to increase the power and ultimately find the genuine pathogenic variant rather than a neutral variant segregating with the family and not the disease. Ten families out of a total fourteen in the cohort did not have a causal mutation in the coding regions of the established SRNS genes and therefore they were studied through segregation analysis. Rare variants segregating with affected members of each family were explored under different model of inheritances to determine if there were

recurrent variation within the same genes across families. Variants in *MUC4* were found across six different families. However, the sample size was limited as there were only ten families and some of them only had two relatives. In order to demonstrate that *MUC4* could be responsible for the SRNS in families, further analysis is require using a control group (Table 11 and Table 12).

In Family A, no shared variants in the coding regions were found in the affected members of the family. Previous studies have identified variants in noncoding sequences to be disease causing, although to date no studies have described noncoding variants responsible for nephrotic syndrome (205). Here, Family A was analysed by parametric linkage analysis which produced promising results showing a region in chromosome 2 that might contain the causal variant for SRNS within the family (Figure 32). However, the region did not reach genome-wide significance suggesting that the dominant parametric model used by MERLIN was very strict (Table 4). Additionally, the variant causing SRNS in the family might not be fully penetrant and it could be present in unaffected relatives which could lead to a low LOD score. Further analysis must be conducted in order to narrow down the region by recruiting more family members.

Seven families were also studied through nonparametric linkage analysis. Two regions in chromosome 2 and chromosome 7 were identified with suggestive evidence of linkage but further evidence would be required to confirm linkage (Figure 33). For both peaks the family contributing the most to the signals was Family A, which is the biggest family in the whole cohort (Table 15 and Table 16). Whilst this was a good exploratory technique, more families with larger pedigrees will be required to narrow down the regions to confirm findings.

For the case control analysis, the sequencing data was assessed by carefully chosen methods and very stringent quality control steps. A joint variant calling was performed to avoid any systematic read depth bias between cases and controls. In this analysis, only European individuals (major ethnicity from the cohort) were studied to avoid different distribution in rare variants caused by populations going through bottle neck events. Therefore, cases and controls were appropriately matched in terms of depth sequencing and ethnic composition (Figure 35). This step is crucial to assure that any statistically significant finding resulting from the analysis is not due to technical differences between datasets. Despite the string quality control steps, the effects of systematic biases were observed in the presence of false associations specifically in the first burden test (without filtering by CADD score). This analysis showed the difficulties associated with combining sequencing data from different technologies. Because rare variants are not frequently observed in the population the effects of different sequencing depth across different platforms are very difficult to correct (Figure 36). Previous studies have reported that WES provides non-uniform coverage when compared to WGS. Additionally, differences in the exome capture technology between the cases and controls could also lead to variation in sensitivity and specificity during variant calling. On the other hand, because of the strict filters some true causal variants might be excluded from the analysis. Two burden tests were performed aggregating variants within a gene with slightly different approaches. In the first burden test, variants were divided by model of inheritance and categorised into three groups: alteration, truncation and synonym. Unexpectedly, some of the genes with the highest association were from the synonymous group. Because most of the synonymous variants are neutral and are not responsible for Mendelian diseases, the synonymous group in this study was used as a negative control. Therefore, it was not

expected to find a higher number of synonymous variants within a gene in cases versus controls. This suggests that most of these signals are artefacts due to sequencing technology despite the strict quality control steps (Table 17) (Figure 38.A). In the second burden test, variants were aggregated by model of inheritance and filtered by CADD score. The results suggested that the pathogenicity filter could overcome the issue experienced in the previous test and most of the artefacts were not replicated. (Table 18) (Figure 38.B)

Overall, this chapter described some preliminary genes that might be associated with SRNS in recessive and dominant model, but further analysis and validation must be carried out. Additionally, despite the fact that sequencing of healthy population individuals has been conducted for European populations, appropriate control sequences for other ethnicities are still lacking. This cohort includes 59 South Asians cases (Figure 20) which could be studied by association analysis with the appropriate controls.

Chapter 6 – Common variant predisposition to SRNS

6.1 Introduction

Both rare and common genetic variation are known to contribute to risk of SRNS, although majority of variants associated with this disease are very infrequent in the standard population. Over the years, several linkage studies have successfully identified genomic regions implicated in SRNS and these types of methodologies are most likely to detect rare and highly penetrant whose frequency in the population remains low due to selection. This is because linkage studies normally lack the power to discover common variants of small effect sizes as the penetrance of individual variants will not be sufficient to observe cosegreagation. Additionally, due to the low prevalence of the disease, most of the SRNS study cohorts available for research are limited in size. Therefore, the impact of common genetic variation on SRNS risk and the contribution to the pathology remains poorly understood. The only common variants well characterised in SRNS are the alleles G1 and G2 in *APOL1* (84). These risk alleles are a rare example of common coding variants that lead to amino acid changes that have large effects on disease susceptibility (up to 30-fold increased risk of renal disease). However, G1 and G2 are only relevant in patients of African descent and no risk alleles have been identified in other ethnicities (162).

In this chapter, to examine if common genetic variation contributes to the risk of disease development in European individuals whose nephrotic disease is unexplained by rare genetic variation in genes previously described in monogenic SRNS, a genome-wide association study was performed. GWAS tests common genetic variants across the whole genome systematically to detect genetic variants that are associated

to diseases and quantitative traits using a logistic regression model. This technique compares the frequencies of each genotype in samples with the phenotype of interest, with the frequencies found in unaffected controls. Thus, controls must be ascertained from a population with similar genomic ancestry to cases in order to avoid false positives. Furthermore, the findings from the GWAS analysis were replicated in an independent cohort (case and control samples) to ensure that the signals found are not specific to a group of cases and can be applied to SRNS patients of European descent.

6.2 Cohort description

For the common variation analysis in this chapter, only the individuals whose DNA had undergone WGS were selected to systematically interrogate their entire genome. Therefore, the SRNS cohort sequenced as part of the *100,000 Genomes Project Rare Diseases Pilot* was studied (section 2.1.1) (Figure 11). Of the total 277 individuals that were whole genome sequenced in the SRNS domain, only 177 had European descent. The cohort comprised 161 sporadic cases, 5 familial cases (probands) and 10 family members. Only a single affected individual from each family was selected for the analysis (just including probands). Any cases with rare putative causal variants that explained their phenotype were excluded. After final phenotypic and renal biopsy review, a further sample was excluded because the biopsy in fact showed likely IgA nephropathy (Berger's disease) and not SRNS (which is characterised by FSGS or MCD without IgA deposition), resulting in a case cohort totalling 159 individuals. Controls were also selected from the same NIHR BioResource project (section 2.1.2) (Table 2), across 13 domains that did not have a renal phenotype, cancer or large effect associations reported for common variants, resulting in a total of 4,405 controls.

Among the 159 SRNS cases, 57.9% were male and were 42.1% female. Clinical phenotypes were paediatric (50.3%) and adult (49.7%) onset and the majority were sporadic cases. The overall mean age of onset was 26.7 years. The onset of SRNS followed a bimodal distribution, characterised by a first peak extending from birth to the second decade and a second peak from the fifth to sixth decades as the one described in section 3.4.1 (Figure 23). Additionally, cases were divided into two clinical subgroups depending on their initial response to steroids: primary SRNS (51.6%) and secondary SRNS (48.4%). Patients had a histological renal biopsy diagnosis, 58.5% were FSGS and 41.5% were MCD. Clinical outcomes were: chronic kidney disease, stable on medication (with or without proteinuria and normal albumin levels) and recovery, defined as no relapse for five years (Table 20).

Table 20. Cohort description of the European WGS samples. Differential profiling in patients with nephrotic syndrome (n=159).

Category	N of patients (%)
Gender	
Males	92 (57.9)
Females	67 (42.1)
Onset	
Paediatric	80 (50.3)
Adult	79 (49.7)
Type	
Primary SRNS	82 (51.6)
Secondary SRNS	77 (48.4)
Histology	
FSGS	93 (58.5)
MCD	66 (41.5)
Outcome	
CKD and transplant	47 (29.6)
Stable on medication	83 (64.2)
Recovery	29 (18.2)

6.3 Data quality control

Most of genome-wide association studies described in the literature analyse genotype data that was generated using SNP arrays combined with imputation. In this analysis whole-genome sequencing data from cases and controls was used instead. Thus, the protocols used for genotype calling and quality control were different to the ones described in the standard imputation-based sequencing technologies. The pipeline used to process the WGS samples is described in detailed in the section 2.6. Variants were called independently across samples and all genotypes were stored together in a multi-sample VCF file performed by the HPC team at the University of Cambridge (102).

In order to run the common variant genome-wide association study, variants with a minor allele frequency higher than 0.05 were selected. In addition, variants were excluded of the analysis if they had a call rate <0.90 or deviated from Hardy-Weinberg equilibrium (P <1 x 10-6). Additionally, a genotype quality (GQ) threshold of 30 and depth (DP) threshold of 20 were set per genotype. All sites with a filter flag other than PASS were removed to ensure selection of high-quality genotypes. Any previously unidentified ancestry outliers were detected with principal component analysis and excluded from downstream analysis (Figure 39). Controls were appropriately matched to cases in terms of genomic ancestry, with 4405 individuals of European ethnicity selected across 11 rare diseases domains and 2 domains with apparently healthy individuals (section 2.1.1) (Table 21). Genetic relatedness was assessed using a subset of high-quality independent common variants and a maximum unrelated set of samples was generated (PI HAT< 0.09375, 2nd degree relationships or closer were

removed). Following quality control, a total of 3,944,568 genotyped variants were retained and used in the association analyses.

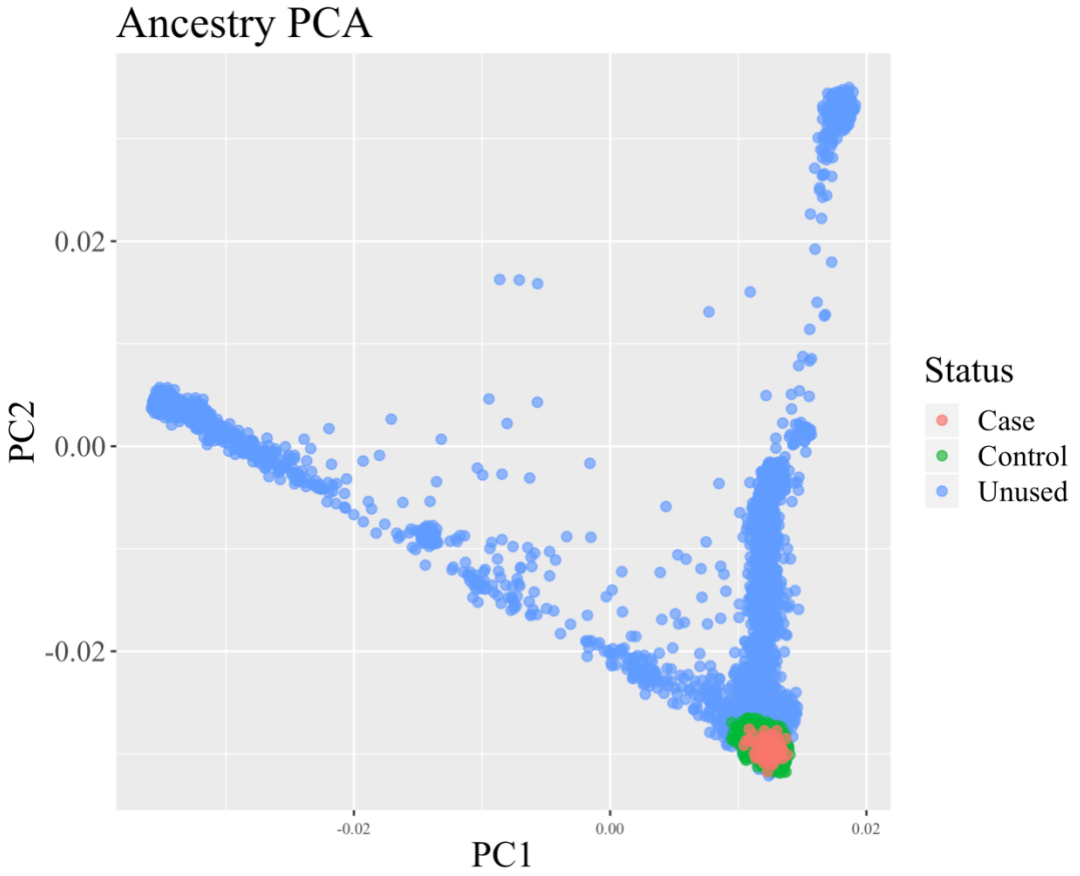


Figure 39. Principal component analysis of the cohort comprising European SRNS cases (n=159) and controls (n=4405). The plot shows the projection of individuals onto the first two principal components of genetic variation. Cases are shown in red dots and controls in green dots. The unused dots in blue are other samples that were not used in the analysis and individuals within the 1000 Genomes Phase 3 data. PCA coordinates were obtained from PC-AiR (206).

Table 21. Number of participants in each study domain from the NIHR BioResource after quality control steps. 18 study domains part of the pilot study for the 100,000 Genomes Project. After post-QC only unreleated participants were selected

NIHR BioResource Domains	Total	European	Post-QC
100,000 Genomes Project - Rare diseases pilot (GEL)	4889	1702	1083
Pulmonary Arterial Hypertension (PAH)	1216	1021	-
Primary Immunodeficiency Diseases (PID)	1430	1126	-
Bronchopulmonary Dysplasia (BPD)	1206	1010	848
Extreme Red Cell Traits (UK Bionbank)	766	750	750
Inherited Retinal Dystrophy (IRD)	736	443	436
Neurological and Developmental Disorders (NDD)	688	448	364
Multiple Primary Malignant Tumours (MPMT)	633	558	-
Intrahepatic Cholestasis of Pregnancy (ICP)	306	218	217
Steroid Resistant Nephrotic Syndrome (SRNS)	277	183	159
Hypertrophic Cardiomyopathy (HCM)	269	240	227
Stem Cell and Myeloid Disorders (SMD)	267	137	86
Cerebral Small Vessel Disease (CSVD)	260	241	132
Neuropathic Pain Disorder (NPD)	210	152	141
Membranoproliferative Glomerulonephritis (PMG)	195	166	-
Tenchnical Controls (CNTRL)	73	51	51
Leber Hereditary Optic Neuropathy (LHON)	72	68	52
Ehlers-Danlos Syndrome (EDS)	23	22	18

6.4 Genome-wide association study

A genome-wide association study was performed to assess if common genetic variation contributes to the risk of SRNS. Therefore, as mentioned in the previous sections, the study focused on individuals that did not have rare genetic pathogenic variants in the coding regions of genes previously described to be causal in SRNS. The case-control analysis was carried out in 159 cases and 4405 controls at 3,944,568 common variants with a logistic Wald association test (EPACTS) (section 2.9.2.3) (Figure 40). The model included the first four principal components as covariates. A summary statistic was reported where each locus was systematically tested for disease association. Examination of the distribution of the observed p-values in the quantile-quantile (QQ) plot and the lambda value ($\lambda GC=1.00$) indicated that population

stratification and genomic inflation were controlled by the stringent QC criteria of genotypes (Figure 41). Four genetic variants in strong linkage disequilibrium reached genome-wide significant association at the MHC region on chromosome 6p21.3 (Table 22). The variant with the strongest evidence of association 6-32584625-A-G (p-value=7.41E-09, odds ratio [OR]=2.321, 95% CI, 1.744 to 3.088), was located in the intergenic region between *HLA-DRB1* and *HLA-DQA1* (Figure 42). The allele was observed at a frequency of 0.8032 in cases compared with 0.6379 in controls. The frequency in the control group was consistent with the frequency of this allele observed in gnomAD (AF=0.6198) for Europeans (non-Finnish). Outside the HLA region, some suggestive signals located in different genomic regions were observed in chromosome 4, 9, 12 and 15 but none of them reached the multiple testing genome wide significance threshold and would require an increase in cases to explore their significance further.

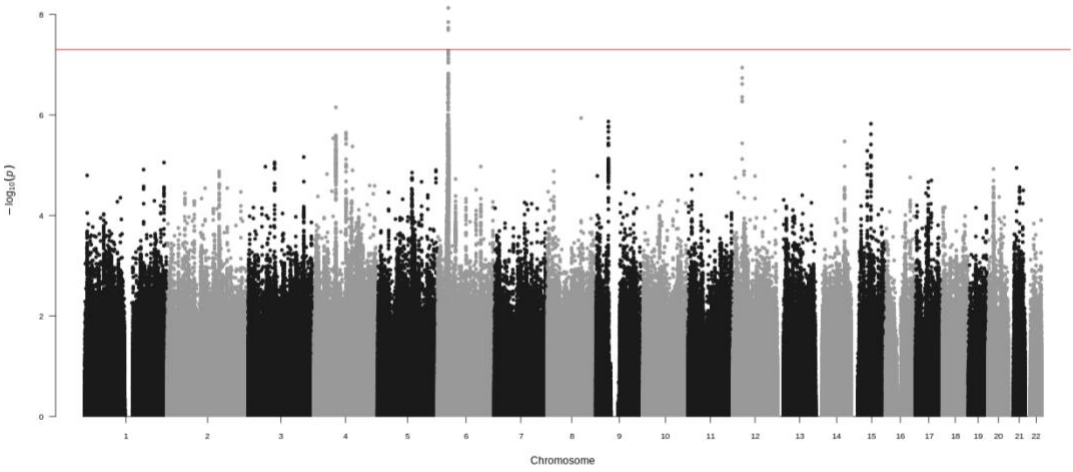


Figure 40 Summary of genome-wide association study results by chromosome. Association with SRNS was determined comparing unrelated European cases (n=159) with controls (n=4405) for 3,944,568 SNPs. Each SNP was tested for association by logistic Wald association test. The x axis represents the position on each chromosome from the p terminus to the q terminus, and the y axis shows the P values on a logarithmic scale. The red line indicates the threshold for genome-wide significance (P=5x10-8).

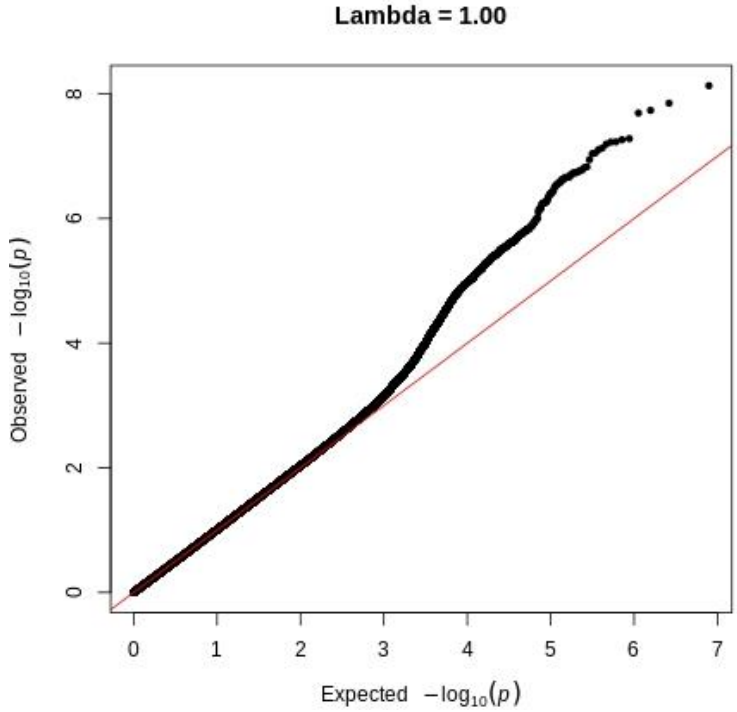


Figure 41. A quantile-quantile plot of GWAS summary statistics (lambda= 1.00). The plot displays the quantile distribution of observed p-value. The x and y axes show the expected and observed logistic regression -log10(P). The red line shows x=y

Table 22. Results of the genome-wide significant analysis comparing SRNS cases with controls (P < 5 x 10-8). The SNPs were tested for association by logistic Wald association test. Four variants reached genome-wide significance and their minor allele frequencies for cases and controls and odd ratios with 95% cofidence intervals are shown: chromosome (Chr), reference allele (Ref), alteration allele (Alt), allele frequency (AF) in cases and controls, allele frequency for Europeans reported in gnomAD and odds ratio (OR) with confidence intervals (CI).

SNP ID	Chr	Position	Ref	Alt	AF Case	AF Control	AF gnomAD	P-value	OR (95% CI)
rs3129758	6	32584625	A	G	0.803	0.637	0.62	7.41E-09	2.321 (1.744-3.088)
rs9271269	6	32580820	G	T	0.589	0.42	0.424	1.84E-08	1.94 (1.54-2.445)
rs9271376	6	32587113	G	A	0.899	0.751	0.731	1.42E-08	2.904 (2.009-4.198)
rs9274660	6	32636434	A	G	0.771	0.61	0.585	2.04E-08	2.2 (1.67-2.898)

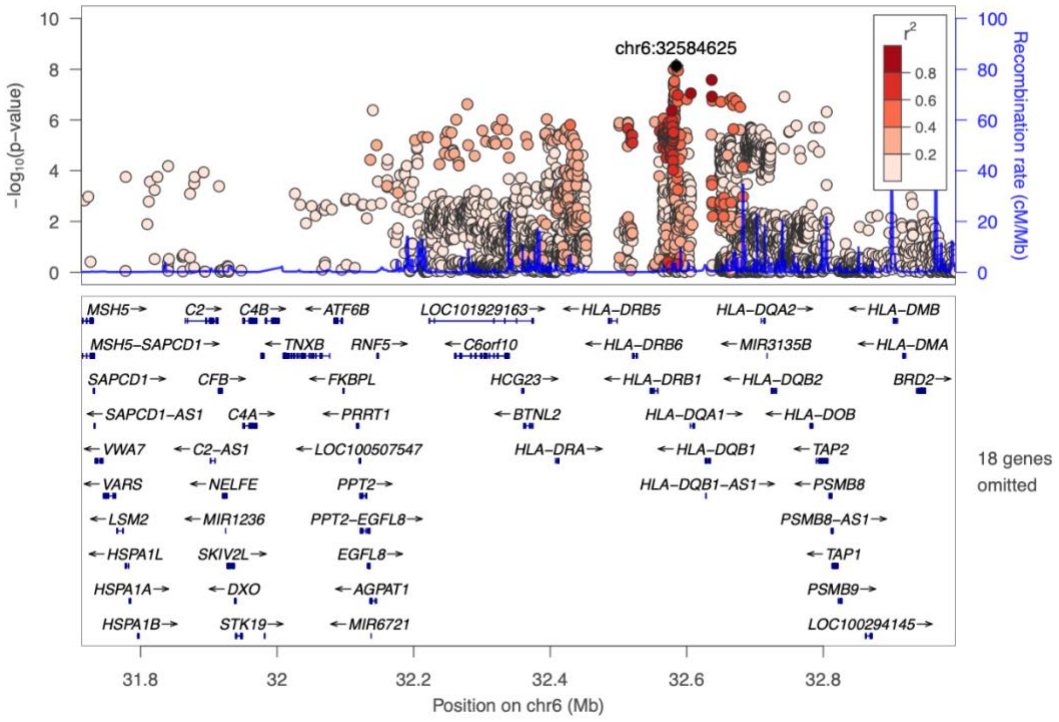


Figure 42. Locuszoom association plot of the GWAS results. The -log10 of the p values are plotted against their physical chromosomal position. Each point represents an analysed SNP by logistic regression, with the strongest association SNP represented by a black diamond.

6.5 Association analysis of classical HLA alleles

To further investigate the allelic basis of the observed GWAS results, the MHC locus was examined in detailed. Because this region is very polymorphic, majority of processing pipelines fail to correctly map the HLA genes since many of the HLA genome sequences present in the population are naturally divergent from the reference genome sequence. Thus, other alignment strategies are needed to perform accurate genotype calls within the region (section 2.8).

6.5.1 HLA genotypes from WES and WGS

As described in the section 2.8.1, two different methodologies were used to call genotypes for HLA alleles in a set of samples that underwent both, WES and WGS. The level of concordance across sequencing platforms (WES and WGS) was compared using both strategies. The tool HLA-Genotyper was chosen because of the high accuracy of genotype calls. Therefore, HLA-Genotyper was used to determine classical HLA alleles from sequencing reads spanning for MHC Class I (*HLA-A, -B* and -*C*) and MHC Class II *(HLA-DR* and *HLA-DQ*) for each sample in the cohort. To further improve the accuracy of the HLA genotype calls ethnicity information was included. There was a total of 270 different HLA alleles for Europeans based on transplant registry frequencies and supplemented HLA alleles found in the 1KGP. HLA alleles were represented by four-digit codes, where the first two represent related group of similar alleles and the third and fourth represent specific proteins with different amino-acid sequences.

6.5.2 HLA dosage-based analysis

A case-control association test was performed on the estimated HLA alleles using a logistic regression model explained in section 2.10.3. The first four genetic principal components were included as covariates to control for population structure effects. The HLA allele with the strongest evidence of association was *HLA-DQA1*01:02* (p-value= 1.38E-07, odds ratio [OR]=0.322, 95% CI, 0.211 to 0.491) (Table 23). The protective allele *HLA-DQA1*01:02* had a frequency of 0.072 in the case cohort compared with 0.203 in controls, the frequency in the control cohort was consistent with the frequency of this allele observed in HapMap (AF=0.238) for Europeans. The SNP with the strongest evidence association from the previous genome-wide

association study was in linkage disequilibrium with *HLA-DQA1*01:02* (r^2=0.66, D'=0.91). This result supported *HLA-DQA1*01:02* as conferring a protective influence against SRNS disease risk.

Table 23. HLA allele association test results. HLA alleles from SRNS cases and controls were compared using a logistic regression test. The table contains information about the allele frequency (AF), odds ratio (OR) and confidence interval (CI) of the alleles with the strongest evidence of association.

HLA Allele	AF Cases	AF Controls	OR (95% CI)	P-Value
*DQA1*01:02*	0.072	0.203	0.32 (0.209-0.487)	1.16E-07
*DRB1*03:01*	0.3	0.172	1.66 (1.353-2.036)	1.18E-06
*DQB1*02:01*	0.315	0.194	1.734 (1.388-2.166)	1.21E-06
*DRB1*15:01*	0.05	0.182	0.336 (0.213-0.53)	2.68E-06
*B*08:01*	0.228	0.135	1.917 (1.458-2.519)	3.06E-06
*DQB1*06:02*	0.056	0.159	0.36 (0.228-0.57)	1.26E-05

6.5.3 Conditional analysis

The genetic variants in close proximity with the associated variant at the associated MHC locus are likely to have inflated values in the results from the association test due to the linkage disequilibrium of the region. Thus, not all the variants that reached genome wide or suggestive significance represent causal associations as they may represent a cluster of variants in LD tagging the same causal variant. In order to test if the HLA allele with the strongest evidence of association was responsible for the signal found in the GWAS results, a conditional analysis was conducted. A GWAS was performed using the estimated alleles from *HLA-DQA1*01:02* as covariates. The results shown that the association of rs3129758 (the SNP with the highest evidence association from the first GWAS) was substantially reduced after conditioning on *HLA-DQA1*01:02* (Figure 43).

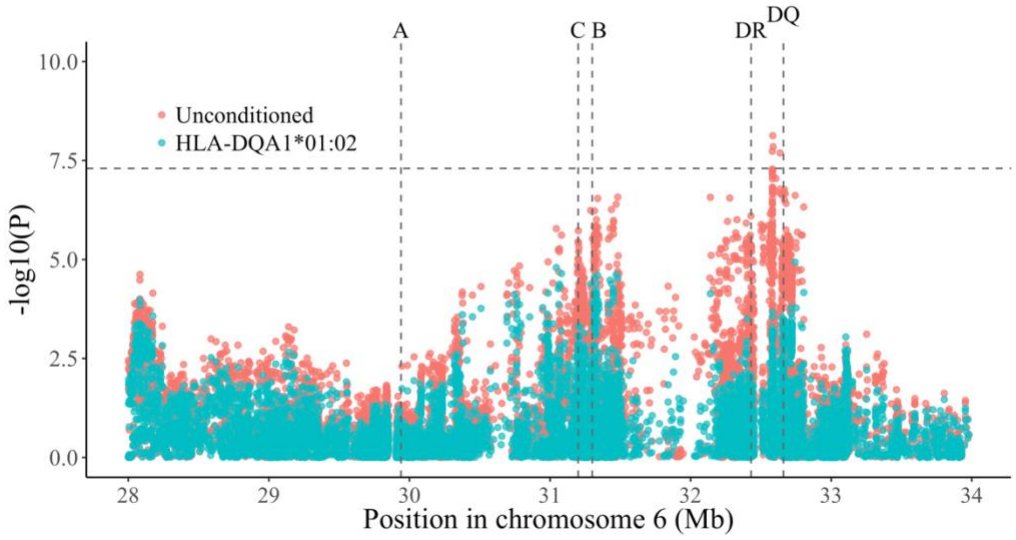


Figure 43. Conditional analysis. Regional association plot showing the results of the unconditioned test (red) and after controlling for *HLA-DQA1*01:02* (green). The -log10 of the p values are plotted against their physical position in the MHC genomic region on chromosome 6. Each point represents an analysed SNP by logistic regression. Conditioning on HLA-DQA1*01:02 left no genome-wide significant signal. The dashed horizontal line indicates the threshold for genome-wide significance (P=5x10-8). The dashed vertical lines indicate the positions of the classical HLA genes *(HLA-A, HLA-B, HLA-C, HLA-DR* and *HLA-DQ*).

6.6 Replication

In order to validate the findings of the GWAS study, a replication analysis using an independent cohort was performed. Reproducibility is crucial to demonstrate that the association found in the MHC locus with SRNS was not by chance or an artefact due to sequencing biases from a specific cohort. Additionally, evidence from other datasets can improve the estimates of the effect sizes of the risk being studied. Thus, genotypes of classical HLA alleles were determined in the SRNS WES cases (section 2.1.1) and the control from the 1958 British Birth Cohort (section 2.1.2). As shown in Figure 11, the proportion of SRNS subphenotypes was different between the discovery and replication cohort due to distinct qualification criteria during the recruitment of cases for the WES and WGS project.

After QC, there were a total of 101 SRNS cases and 936 population controls in the replication cohort. Exome sequencing reads were mapped to gene *HLA-DQA1*, to estimate genotypes of alleles at four-digit resolution in cases and controls. Association analysis was then performed on the estimated *HLA-DQA1* alleles using a logistic regression model including the first four principal components as covariates. The protective effect of *HLA-DQA1*01:02* on SRNS disease risk was confirmed in the replication cohort (p-value=0.038, odds ratio [OR]=0.640, 95% CI, 0.41 to 0.97). The allele *HLA-DQA1*01:02* had also a reduced frequency of 0.138 in cases compared with 0.201 in controls. The direction of this association was consistent between the discovery and replication cohort. The point estimate was larger in the discovery cohort but the confidence intervals did overlap between the discovery and replication cohort (Figure 44).

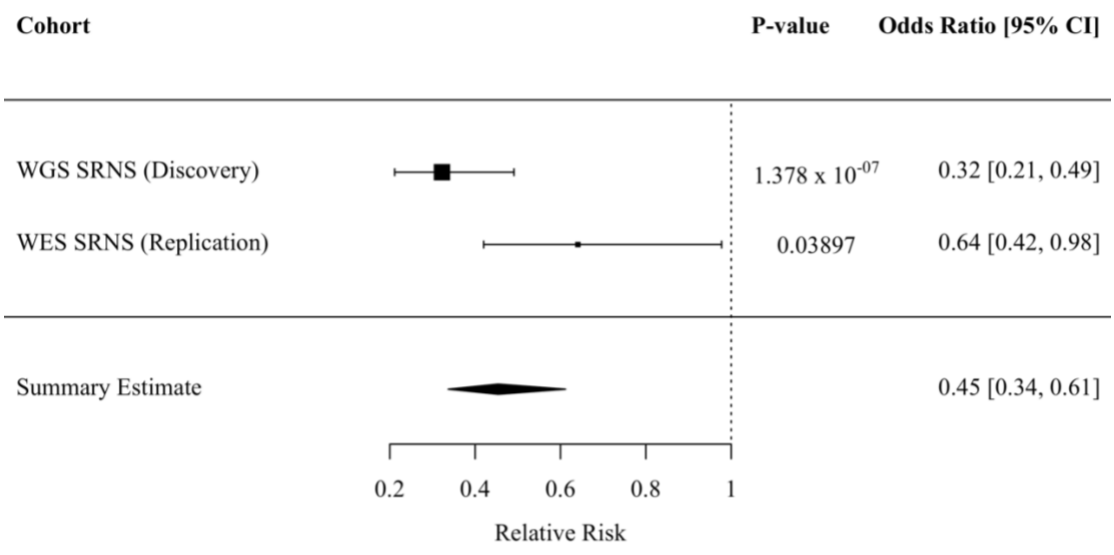


Figure 44. Effect size estimates for HLA-DQA1*01:02 on risk of SRNS. Effect size estimates for HLA-DQA1*01:02 in the discovery cohort determined from WGS data and in the replication cohort determined from WES data. Both cohorts were combinations of European case/control populations.

6.7 The effect of HLA-DQA1*01:02 across SRNS subphenotypes

Since the SRNS phenotype is relatively heterogeneous, the study included subgroups of cases with different ages of onset and disease severity. Consequently, additional linear regression analysis was performed (described in Section 2.10.4) to ascertain any potential association between *HLA-DQA1*01:02* and phenotypic variables such as gender, primary or secondary SRNS, histology type (FSGS or MCD) and age-of-onset. The test was corrected for population structure including the first four principal components as covariates. This showed that none of the variables were significant associated with *HLA-DQA1*01:02*: gender (p-value=0.871, Estimate=-0.730 and Std.error=1.849), SRNS type (p-value=0.541, Estimate=1.84 and Std. error=1.81), histology type (p-value=0.521, Estimate=0.424 and Std. error=0.041) and age-of-onset (p-value=0.132, Estimate=-94.084 and Std.error=89.931).

The effect of *HLA-DQA1*01:02* was also explored in the familial cases and also in individuals where a causal mutation or mutations were found for monogenic forms of SRNS. In the WGS cohort there were only 3 European families. Although Family A has been described as a WES family in chapter 3, three relatives from the family underwent WGS too. The protective allele *HLA-DQA1*01:02*, did not seem to be more prevalent in unaffected members across the three families, as there were three affected cases that were heterozygous for the protective allele (Table 24). Therefore, *HLA-DQA1*01:02* did not cosegregate with the disease. Furthermore, of 7 European cases whose SRNS was explained by a rare genetic mutation in one of the established genes, 3 were heterozygous for *HLA-DQA1*01:02* (Table 25). Due to the limited

number of both groups, familial cases and 'solved' cases, frequencies of the allele *HLA- DQA1*01:02* could not be compare with the 159 cases studied in the GWAS.

Table 24. HLA-DQA1*01:02 genotypes in familial cases.

FAMILY	ID	STATUS	HLA-DQA1*01:02 allele
3	S013341	Affected	0/0
	S013718	Unaffected	1/1
	S013754	Unaffected	0/0
4	S017250	Affected	0/0
	S017274	Affected	0/1
	S017262	Unaffected	0/0
A	S006337	Affected	0/0
	S012009	Affected	0/1
	S012021	Affected	0/1

Table 25. HLA-DQA1*01:02 genotypes in cases with a mutation in one of the establsihed SRNS genes.

ID	GENE	HLA-DQA1*01:02 allele
S012696	*TRPC6*	0/0
S013421	*CRB2*	0/0
S014415	*NPHS1*	0/1
S014334	*NPHS2*	0/1
S014195	*WT1*	0/0
S014194	*WT1*	0/1
S013282	*NUP93*	0/0

6.8 Investigation of other putative SRNS genetic association signals

Genome-wide association studies have been successful in identifying many loci associated with complex traits or diseases. However, the causal mechanisms underlying these associations are not always clear and require further analysis. Thus, colocalisation approaches have been used to search for the causal mechanisms responsible for a trait integrating protein expression data from tissues of interest. Colocalisation is a statistical methodology that combines the GWAS summary statistics with gene expression data to determine if the same underlying variant is responsible for both disease risks (156).

Here, the GWAS results (excluding the MHC region due to its high LD) were tested to see if the suggestive peaks in chromosomes 4, 9, 12 and 15 were colocalising with kidney gene expression signals. The kidney expression data was obtained from public resources, which contain genetic variants that explain a portion of variance in expression of a set of genes, also known as expression quantitative trait loci (eQTLs). The methodology and data used is described in section 2.10.5. A genetic colocalization was performed using a set of variants that overlapped (within the same region of the genome) between both datasets under the hypothesis that both traits are associated and share a single causal variant. None of the variants reached statistically significant association. Thus, there was no overlap between the signals of both datasets.

6.9 Discussion

Through association analysis of 159 SRNS cases and 4,405 controls of European ancestry using whole-genome sequencing data, a genome-wide significant locus in the

MHC region between *HLA-DRB1* and *HLA-DQA1* was identified (Figure 40). The HLA genes are highly polymorphic and responsible for the adaptive immune reposes together with the innate immune responses to infectious. Many complex diseases including autoimmune, infectious and inflammatory diseases as well as cancer have been associated with the MHC complex. Genotypes of HLA alleles were estimated through realignment of unmapped reads from WGS data against a comprehensive European reference panel. An association analysis showed the MHC class II allele, *HLA-DQA1*01:02*, to be protective for SRNS and had the highest evidence of association (p-value= 1.38E-07, odds ratio [OR]=0.322, 95% CI, 0.211 to 0.491) (Table 23). This association was also replicated in an independent European SRNS cohort that included a mixture of primary and secondary steroid resistant nephrotic syndrome patients using WES data.

The effect of *HLA-DQA1*01:02* across SRNS subphenotypes was also explored. There were not significant differences in the frequency of *HLA-DQA1*01:02* in clinical subgroups (disease type, histology, age of onset) of the SRNS cases. However, the sample size was very limited, as there were only 159 SRNS cases. The cosegregation of the allele was also explored in familial cases and cases whose nephrotic syndrome was explained by a causal mutation in one of the established genes. Again, sample size was limiting as only three families were available. Therefore, the conclusions made on the possible effect of this protective allele in one of the subphenotypes of the disease were very restricted and a larger sample size is required.

Although the role of the immune system in the SRNS aetiology was suspected because of the efficacy of immunosuppressive treatments to alleviate symptoms of affected

individuals, this finding confirms the link between the immune system and SRNS and might offer new leads into the pathogenesis of the disease. HLA class II genes are expressed in antigen presenting cells such as B cells and are specifically responsible for the presentation of self or foreign peptides to $CD4^+$ T cells, to initiate an adaptive immune response. The lower frequency of *HLA-DQA1*01:02* in SRNS cases in comparison with controls might suggest that the antigen presentation is disrupted or impaired affecting the adaptive immune response.

Furthermore, many other studies have identified a significant genome wide association with a locus in *HLA-DQA1* and SSNS cohorts of European (*HLA-DQA1*01*, p-value=1.90E-31, odds ratio [OR]=0.36, 95% CI, 0.30 to 0.43) and South Asian origin (rs1129740, p-value=1.18E-6, odds ratio [OR]=2.11). Traditionally, SSNS and SRNS have been considered two independent diseases marked by the response to treatment. Thus, SRNS has normally been considered a disease caused by gene defects in proteins important for the function of the glomerular filtration barrier, whereas SSNS has been described as an immune disease since the genetic architecture remains poorly understood and the only genetic findings described were polygenic risks in *HLA-DQA1* and *HLA-DQB1*. Therefore, this association with *HLA-DQA1* is not exclusive of SSNS (56), as primary and secondary SRNS are also associated with the gene. This finding suggests commonality in the molecular basis between SSNS and SRNS in Europeans, implying that molecular diagnosis might be more informative than clinical labels.

Chapter 7 – General discussion

Nephrotic syndrome has been studied over hundreds of years. There are descriptions of the disease since the golden age of Athens made by Hippocrates: “when bubbles settle on the surface of the urine, it indicates a disease of the kidney, and that the disease will be protracted” (207, 208). It was not until 1827 when Sir Richard Bright connected the classical symptoms of nephrotic syndrome not only with a disease of the kidney but with excessive amounts of protein lost into the urine and this was the underlying cause (208).

The first genetic discoveries responsible for SRNS were not identified until 1991, when mutations in *WT1* were identified in affected individuals with Denys-Drash syndrome characterised by nephrotic syndrome phenotype (66). To date, advances in massive parallel sequencing technologies have resulted in sequencing cost declining and consequently whole exome and whole genome sequencing has become widely used for the study of Mendelian disorders including SRNS. Therefore, this revolution in the field of genetics have facilitated the discovery of more than sixty genes that are mutated in nephrotic syndrome. Nevertheless, understanding of SRNS disease aetiology remains unclear and it is likely that both genetic and environmental factors are involved in the course of the disorder. Consequently, the treatments available alleviate symptoms and complications but seldom results in a cure.

This thesis investigated the molecular genetics of idiopathic nephrotic syndrome in children and adults, specifically SRNS forms, in one of the largest cohorts that has undergone whole exome and whole genome sequencing in the UK. Detailed phenotypic data was available including type of SRNS, age of onset, histology, clinical

outcomes and medications used as explained in chapter 3. The cohort comprised of paediatric and adult SRNS cases drawn from renal populations found in clinics across the UK and it was majority formed by sporadic cases of European ethnicity. The onset of SRNS followed a bimodal distribution, characterised by a first peak extending from birth to the second decade and a second peak from the fifth to sixth decades (Figure 23). This was a novel finding as classically SRNS is considered a disease of early life. Most of individuals had early age of onset and the most common long-term clinical outcome was chronic kidney disease or transplant.

Three main objectives were pursued to identify causal genetic variation and to further understand aspects of the disease biology that remain poorly understood. Firstly, rare genetic variation in the coding regions of previously associated SRNS genes was evaluated, then rare genetic variation in exome of novel genes was also explored and finally, common genetic variation within the entire genome of affected individuals was also studied.

7.1 Summary

In chapter 4, the coding regions of 67 genes previously associated with SRNS were screened in a cohort of 422 individuals from different ethnicity groups. Samples were recruited from two distinct projects using different sequencing technologies, then merged and processed using the same WES data pipeline to minimise systematic sequencing errors. 48 patients had at least one putative mutation in a known SRNS gene that explained the phenotype (Table 7) (Table 8). The pathogenic variants detected were rare and highly penetrant. Despite the rare status, half of the diagnosed cases had a mutation that was previously reported in familial or sporadic SRNS cases by other studies (Table 7). By contrast, the other half of cases had at least one novel

variant where detailed phenotype analysis and bioinformatic prediction tools were crucial to assign likely pathogenicity (Table 8).

Of the total 67 known genes, only 13 were found mutated in the cohort*: NPHS2, NPHS1, MAGI2, WT1, LM1XB, NUP93, CRB2, COL4A3, MYO1E, NUP107, LAMB2, LCAT* and *TRPC6* (Figure 25). Most of the pathogenic variants were identified in *NPHS2* and *NPHS1* which is similar to what it was reported by other studies, partially explained by some bias towards childhood onset SRNS during recruitment. However, it was surprising not to detect other genes connected with childhood and adult SRNS that are considered "major causes" such as *INF2* considering the number of adults recruited to the cohort. *APOL1* risk alleles were present as expected in cases of African descent and not in other ethnicities.

The variety of mutations found in each of the different genes underpinned the considerable heterogeneity present in this disease. Nevertheless, majority of known genes were not mutated and therefore responsible for disease in this cohort and the incidence of mutations in the established SRNS genes was lower than previously reported, due to a tendency to recruit sporadic cases. This may suggest that some of the variants or genes that have been previously described in the literature might not be pathogenic for SRNS, especially those detected through study of isolated highly interbred families. Improvements in control databases now allow rejection of either genes/variants as non-causal or randomly associated. This has been an issue in previous studies including those that have flagged erroneous classifications of pathogenicity in clinical databases, in particular when candidate variants and allele frequency could not be studied as comprehensively using population databases such as gnomAD.

The families sequenced within the cohort were also screened to find potential causal mutations by extracting variants in the 67 established SRNS genes shared among the affected individuals that were not present in the unaffected. Furthermore, depending on the established mode of inheritance of each gene, dominant and recessive inheritance models were applied. Of a total 14 families, 4 families (Family 5, Family B, Family C and Family G) had a pathogenic variant in one of the established SRNS genes that explained their phenotype and fitted the family data (Figure 26, figure 28-30).

Overall, the screening of the established SRNS genes collectively explained only ~11% of the cases. Since the percentage of diagnosed cases was lower in comparison with other SRNS cohorts, this could be explained by differences in recruitment as majority of affected individuals in the cohort were sporadic cases with heterogonous age of onset unlike the other studies that focused on paediatric populations with familial disease and very early age of onset. Additionally, some cases were screened for established SRNS genes prior to recruitment, introducing potential recruitment bias towards affected individuals that did not have an obvious monogenic form of the disease. However, since rigorous phenotyping of all cases was possible, this ensured there was no misdiagnosis and this figure was likely to represent an accurate incidence of causal mutations in known SRNS gene within this cohort.

In chapter 5, the remaining affected individuals that were not explained by genetic variation in the known SRNS genes, were explored to identify rare genetic variation in novel genes. A stringent variant filtering protocol was performed in combination with two main methodologies, family-based analyses (segregation analysis, parametric linkage analysis and nonparametric linkage analysis) and a case control

association test (gene-based burden test). Additionally, this chapter described the extensive QC performed in the cohort to reject technical errors or false positives, as well as the challenges of combining lower coverage whole genome data with whole exome and the importance of performing a joint variant calling to overcome such issues.

The ten families not explained by any causal mutations in the coding regions of the established SRNS genes were studied further using segregation analysis. Rare variants segregating in the affected relatives of each family were explored under dominant and recessive model of inheritance. Rare variants in *MUC4* were recurrent across six families (Table 12). Although *MUC1* mutations have been previously associated with autosomal dominant tubulointerstitial kidney disease and secondary FSGS, the group of genes that encode mucins (*MUC1-9*) are highly polymorphic and have very repetitive sequences. Thus, these genes are poorly mapped and have lots of errors and/or false positives variants. As the sample size in this study was quite limited encompassing only ten families and some were only duos, further analysis with bigger sample sizes would be required to determine any potential causal role of *MUC4* in familial SRNS.

Family A, the biggest sequenced family of the entire cohort, did not have any shared variants in the coding regions segregating with the affected members that were not present in the unaffected. Previous studies have identified disease causing variants in noncoding sequences, however, to date no studies have described noncoding variants responsible for SRNS. To explore this in more detail, Family A was analysed by parametric linkage analysis under the model of a rare dominant and highly penetrant trait. A region in chromosome 2 located between co-ordinates 180,810,180 to

180,835,792 had a LOD score of 2.1154 (Figure 32) (Table 13). The segregation of haplotypes in the region was consistent with linkage and indicated that this could potentially contain the causal variant. However, while the LOD score was suggestive it was not conclusive. Considering the phenotype itself is variably penetrant within this family, this could indicate that the causal variant is not fully penetrant and that the model used for analysis too stringent. Further analysis is required to determine the causality of the region by recruiting more members of the family. Some members have agreed to participate in the analysis, but blood samples have not been obtained yet. Analysis of the extended pedigree would potentially facilitate the discovery of a novel disease gene, since the region did not contain any known SRNS genes. Moreover, it would be the first variant in a non-coding region to be found responsible for SRNS.

A nonparametric or model free linkage analysis was also performed in seven informative families with at least two affected siblings per family. Two regions in chromosome 2 and chromosome 7 were identified with suggestive evidence of linkage (LOD score of 2.811 and 2.319 respectively) (Figure 33). For both peaks the family contributing the most to the signals was Family A, which is the biggest family in the whole cohort. Thus, to narrow down both regions the pedigree sizes should be expanded, and more families should be recruited.

Two gene-based burden tests for rare variants were performed using different filtering criteria in the European individuals of the cohort. Certain limitations of the study design and challenges surrounding using different sequencing technologies are described. For both tests, variants were aggregated within a gene a divided by model of inheritance. In the first, variants were categorised into three groups: alterations, truncations and synonymous, whereas in the second variants were filtered by

pathogenicity as measured by CADD score. The only gene that reached burden significance in both tests was *ZNF780B* under a recessive model. Within this gene two variants were found in compound heterozygosity in 8 cases suggesting that the signal might be an artefact as it is not expected that the same rare variant would be recurrent across unrelated cases. Moreover, both variants were considered 'benign' according to the Polyphen-2. From the burden test results, some genes had suggestive significance that might be associated with SRNS in recessive and/or dominant model. A gene-set analysis was then explored using a hypergeometric test and two different lists of genes associated with kidney disease and kidney expression applied to the burden test results. None of the tests reached statistical significance suggesting there was not an enrichment of mutations in genes highly expressed in podocytes or genes associated with kidney disease. As rare disease analysis is often hampered by sample size, further analysis using a larger cohort is required, and this is planned for future work.

In chapter 6, the aim was to detect common genetic variation with small or medium effect sizes that could be associated with SRNS risk. A genome wide association study was performed using WGS data on 159 SRNS cases and 4,405 controls of European ancestry. Although GWAS is normally used for the study of common complex diseases or traits, this methodology has also been proven to be informative for some rare diseases too such as severe neurodevelopmental disorders (209). Here, a genome-wide significant locus within the MHC region between the class II genes *HLA-DRB1* and *HLA-DQA1* was found to be associated with SRNS (Figure 40). This finding confirmed the importance of the immune system in both children and adults affected by SRNS, suggesting that this disease could fall into the category of an autoimmune disorder. Further analysis of the region through estimation of HLA alleles by

realignment of raw reads that were unmapped showed *HLA-DQA1*01:02* to be protective for SRNS and had the highest evidence of association (p-value= 1.38E-07, odds ratio [OR]=0.322, 95% CI, 0.211 to 0.4917) (Table 23). The association was replicated in an independent cohort that included primary and secondary SRNS individuals that underwent whole exome sequencing (p-value=0.033, odds ratio [OR]=0.637, 95% CI, 0.419 to 0.966). The effect of *HLA-DQA1*01:02* was also explored in the cases subphenotypes although not significant different on the effect of the allele were found in any category (disease type, histology, age of onset) but this may be due to the relatively low numbers of cases available for study.

Previous studies had implicated common genetic variants in the HLA-DR/DQ in SSNS. However, this is the first time an association has been detected for SRNS, including children and adult cases. These findings support the concept that SRNS has an autoimmune basis, and also suggests common molecular genetics with SSNS.

7.2 Genetic heterogeneity and phenotypic variation

The results from this study using a national cohort of children and adults with SRNS confirms that not only is SRNS genetically heterogeneous, but a large proportion of cases are unexplained by classical Mendelian inheritance. While exome screening detected *NPHS2* and *NPHS1* mutations as responsible for majority of the monogenic forms of the disease in the affected individuals of the cohort, there were at least 13 different genes mutated that explained only ~11% of the total cases. This was one of the original drivers to re-sequence some of the original WES cohort by WGS. Interestingly, this did not raise the mutation hit rate which was to some extent unexpected (102). Thus, the genetic heterogeneity has probably affected the identification of causal variants especially those with a modest effect that could be

acting as modifiers or variants with variable penetrance. Furthermore, the identification of a locus in the HLA-DR/DQ region associated with SRNS in European individuals has also proven that the genetics of the disease is more complex than initially thought. This data advocates that undoubtedly, other approaches beyond the classical investigative routes used for monogenic disorders should be considered when studying SRNS.

Another challenge present in the different results chapters was the phenotypic heterogeneity. While, in contrast to many other studies where robust phenotyping is not possible, all cases in this study could be categorised as SRNS and there was high confidence that all misclassified cases where the SRNS was secondary (e.g., due to IgA nephropathy) were excluded. Despite this, affected individuals within the cohort had very different age of onset and symptoms as well as a variety of clinical progression, response to medication and consequently clinical outcomes. In fact, in some of the families described in the cohort, there were very different phenotypes and symptoms within the same family. Family A had some affected relatives with kidney failure that required dialysis whereas others only had mild proteinuria and did not require treatment. At least two families, Family D and Family E, had a mixture of phenotypes with members affected by SRNS and SSNS. Those families potentially shared the same causal mutation among the affected relatives of the family, but their phenotypes are clinically independent, as SRNS and SSNS are often considered two different diseases. This intrafamilial phenotypic variability is not unique of nephrotic syndrome and has been observed in other families affected by Mendelian diseases such as Bardet-Biedl syndrome (210), hypophosphatasia (211) or limb-girdle muscular dystrophies (LGMDs) (212). Additionally, some studies have reported the phenotypic variation present specifically in kidney diseases (213). The role of variants

or genes associated with kidney disease has dramatically changed due to advances in the sequencing technologies and some of the major kidney disease-associated genes are known to be responsible for a broader phenotypic spectrum. Mutations in *COL4A3-5* genes or *DGKE* are responsible for clinically unrelated kidney diseases such as Alport syndrome, FSGS, atypical haemolytic uraemic syndrome and nephrotic syndrome. Therefore, the same genetic background can be responsible for very diverse clinical phenotypes. Findings from this thesis support these observations as the association found with *HLA-DQA1*01:02* and SRNS, which is similar to the loci previously reported for SSNS, suggests common molecular genetics with SSNS (54-56). Accordingly, SRNS and SSNS may share immunological traits within the MHC class II *genes HLA-DR/DQ*. This is supported by the observation that both have an equivalent risk of disease recurrence post-transplant which is attributed to the presence of an immune active circulating factor rather than an intrinsic kidney defect.

Lastly, these findings also raise questions regarding the classification and diagnoses of human diseases in particularly in kidney phenotypes that are not always uniformly expressed such as SRNS. Genetic testing to support clinical grounds such as histology or response to treatment is crucial to ensure a correct diagnosis. Conceivably, in SSNS and SRNS molecular diagnosis might ultimately be more informative than clinical labels allowing correct stratification of the phenotype leading to precision medicine. Currently, some of the established classifications could be redefined under a broader definition that includes genetic evidence and not just clinical observations.

7.3 General technical limitations

One of the main limitations was the modest sample size of the cohort which is an inherent problem common to such studies in rare disease, where networks and

collaborations between different centres are crucial to achieve a successful recruitment of sufficient numbers of affected individuals. A larger sample size could have solved most of the challenges encountered during the different analyses that were performed in this study.

However, to at least in part address the sample size limitation, two independent datasets of SRNS cases were merged and processed using the same pipeline. This was to overcome the analytical challenges of matching data across different sequencing platforms with non-uniform coverage and to avoid the other sources of artefacts and technical variation. This required a joint variant calling and considerable time performing quality control procedures to design and optimise thresholds to filter out any sequencing errors. Previous studies have also reported the considerable efforts required to perform QC assessments that assure high quality sequencing data and also the importance of a joint genotype calling (116, 117).

Despite the improvements in control databases and specifically for European ethnicity, this has not yet been matched for other ethnicities and this remains a significant issue for these types of studies (214). As such, despite the diverse ethnic background of the SRNS cases within the cohort, this had implications on the ability to comprehensively explore the genetic background to SRNS beyond the European group. This was not possible even though SRNS is nominally more common in populations of African and South Asian descent (8, 9). Larger and more diverse control groups are required specially for non-European ethnicities since sequencing of healthy individuals has been conducted mainly for European populations. The lack of control groups for African or Asian populations meant that analysis such as gene-based burden test or the GWAS focused only on European individuals that matched ethnically with our

available controls. The association between HLA-DQA1*01:02 and SRNS should also be explore in other ethnicities. This lack of control groups is not limited only to the present study, but significant improvements are expected in the close future with declining cost of WES and WGS technologies and with the creation of national and international consortiums that promote initiatives to share data.

7.4 Future work

The work presented here has shown that application of whole exome and whole genome sequencing technologies can provide important further knowledge regarding the genetic architecture and pathobiological mechanisms underlying SRNS and nephrotic syndrome in general. The findings also have relevance for other monogenic diseases that in addition, may also present as a common complex trait and show characteristics of a multifactorial disease. As previously mentioned, further studies with larger sample size will be pivotal to significantly improve our knowledge of the disease and identify why some people are prone to it and predict their long-term clinical outcomes. Ongoing national and international initiatives that include multiple recruitment centres as well as robust data-basing of clinical phenotypes to minimise misclassification have been developed to allow identification of sufficient numbers of cases.

Functional experiments are proposed to provide mechanistic evidence for the pathogenic effect of mutations identified in both established and novel genes using animal models and/or human tissue culture models. Furthermore, integration of functional and gene expression information obtained from RNA sequencing of kidney tissue matched with exome or whole genome sequencing would allow to prioritise candidate genes in family-based studies and case control association studies. Evidence

was provided for the involvement of a locus in the HLA region, but how *HLA-DQA1*01:02* allele is associated with SSNS/SRNS risk remain unclear. In view of current treatments for SRNS that target CD-20 antigens on the surface of lymphocytes B (22), it is compelling to consider infection acting as a trigger and a direct/indirect interaction with HLA resulting in a maladaptive immune response.

In conclusion, this work has improved the understanding of the molecular genetic basis of nephrotic syndrome and particularly SRNS forms. Cases with detailed phenotype were crucial to identify correlation between genetics and clinical features that overall could lead to a better diagnosis and treatment for young children and adults with nephrotic syndrome.

References

1. Preuss HG. Basics of renal anatomy and physiology. Clin Lab Med. 1993;13(1):1-11.

2. Gross F, Schaechtelin G, Brunner H, Peters G. The Role of the Renin-Angiotensin System in Blood Pressure Regulation and Kidney Function. Can Med Assoc J. 1964;90:258-62.

3. Salmito FT, de Oliveira Neves FM, Meneses GC, de Almeida Leitao R, Martins AM, Liborio AB. Glycocalyx injury in adults with nephrotic syndrome: Association with endothelial function. Clin Chim Acta. 2015;447:55-8.

4. Kanwar YS, Liu ZZ, Kashihara N, Wallner EI. Current status of the structural and functional basis of glomerular filtration and proteinuria. Semin Nephrol. 1991;11(4):390-413.

5. Edwards A, Daniels BS, Deen WM. Ultrastructural model for size selectivity in glomerular filtration. Am J Physiol. 1999;276(6):F892-902.

6. Schell C, Sabass B, Helmstaedter M, Geist F, Abed A, Yasuda-Yamahara M, et al. ARP3 Controls the Podocyte Architecture at the Kidney Filtration Barrier. Dev Cell. 2018;47(6):741-57 e8.

7. Noone DG, Iijima K, Parekh R. Idiopathic nephrotic syndrome in children. Lancet. 2018;392(10141):61-74.

8. Chanchlani R, Parekh RS. Ethnic Differences in Childhood Nephrotic Syndrome. Front Pediatr. 2016;4:39.

9. Srivastava T, Simon SD, Alon US. High incidence of focal segmental glomerulosclerosis in nephrotic syndrome of childhood. Pediatr Nephrol. 1999;13(1):13-8.

10. Dornan TL, Jenkins S, Cotton RE, Tattersall RB, Burden RP. The nephrotic syndrome at presentation of insulin-dependent diabetes mellitus; cause or coincidence? Diabet Med. 1988;5(4):387-90.

11. Chen D, Hu W. Lupus podocytopathy: a distinct entity of lupus nephritis. J Nephrol. 2018;31(5):629-34.

12. Govers LP, Toka HR, Hariri A, Walsh SB, Bockenhauer D. Mitochondrial DNA mutations in renal disease: an overview. Pediatr Nephrol. 2021;36(1):9-17.

13. Eddy AA, Symons JM. Nephrotic syndrome in childhood. Lancet. 2003;362(9384):629-39.

14. McCloskey O, Maxwell AP. Diagnosis and management of nephrotic syndrome. Practitioner. 2017;261(1801):11-5.

15. Alshami A, Roshan A, Catapang M, Jobsis JJ, Kwok T, Polderman N, et al. Indications for kidney biopsy in idiopathic childhood nephrotic syndrome. Pediatr Nephrol. 2017;32(10):1897-905.

16. Fiorentino M, Bolignano D, Tesar V, Pisano A, Van Biesen W, D'Arrigo G, et al. Renal Biopsy in 2015--From Epidemiology to Evidence-Based Indications. Am J Nephrol. 2016;43(1):1-19.

17. Agarwal SK, Sethi S, Dinda AK. Basics of kidney biopsy: A nephrologist's perspective. Indian J Nephrol. 2013;23(4):243-52.

18. Tryggvason K, Patrakka J, Wartiovaara J. Hereditary proteinuria syndromes and mechanisms of proteinuria. N Engl J Med. 2006;354(13):1387-401.

19. Watanabe A, Feltran LS, Sampson MG. Genetics of Nephrotic Syndrome Presenting in Childhood: Core Curriculum 2019. Am J Kidney Dis. 2019;74(4):549-57.

20. Bierzynska A, Soderquest K, Koziell A. Genes and podocytes - new insights into mechanisms of podocytopathy. Front Endocrinol (Lausanne). 2014;5:226.

21. Colucci M, Corpetti G, Emma F, Vivarelli M. Immunology of idiopathic nephrotic syndrome. Pediatr Nephrol. 2018;33(4):573-84.

22. Iijima K, Sako M, Nozu K, Mori R, Tuchida N, Kamei K, et al. Rituximab for childhood-onset, complicated, frequently relapsing nephrotic syndrome or steroid-dependent nephrotic syndrome: a multicentre, double-blind, randomised, placebo-controlled trial. Lancet. 2014;384(9950):1273-81.

23. Weber S, Tonshoff B. Recurrence of focal-segmental glomerulosclerosis in children after renal transplantation: clinical and genetic aspects. Transplantation. 2005;80(1 Suppl):S128-34.

24. Kuusniemi AM, Qvist E, Sun Y, Patrakka J, Ronnholm K, Karikoski R, et al. Plasma exchange and retransplantation in recurrent nephrosis of patients with congenital nephrotic syndrome of the Finnish type (NPHS1). Transplantation. 2007;83(10):1316-23.

25. Shah L, Hooper DK, Okamura D, Wallace D, Moodalbail D, Gluck C, et al. LDL-apheresis-induced remission of focal segmental glomerulosclerosis recurrence in pediatric renal transplant recipients. Pediatr Nephrol. 2019;34(11):2343-50.

26. Fogo AB. Causes and pathogenesis of focal segmental glomerulosclerosis. Nat Rev Nephrol. 2015;11(2):76-87.

27. Floege J, Amann K. Primary glomerulonephritides. Lancet. 2016;387(10032):2036-48.

28. Nso Roca AP, Pena Carrion A, Benito Gutierrez M, Garcia Meseguer C, Garcia Pose A, Navarro M. Evolutive study of children with diffuse mesangial sclerosis. Pediatr Nephrol. 2009;24(5):1013-9.

29. Nephrotic syndrome in children: a randomized trial comparing two prednisone regimens in steroid-responsive patients who relapse early. Report of the international study of kidney disease in children. J Pediatr. 1979;95(2):239-43.

30. Maas RJ, Deegens JK, Smeets B, Moeller MJ, Wetzels JF. Minimal change disease and idiopathic FSGS: manifestations of the same disease. Nat Rev Nephrol. 2016;12(12):768-76.

31. Lepori N, Zand L, Sethi S, Fernandez-Juarez G, Fervenza FC. Clinical and pathological phenotype of genetic causes of focal segmental glomerulosclerosis in adults. Clin Kidney J. 2018;11(2):179-90.

32. Lumbers ER, Kandasamy Y, Delforce SJ, Boyce AC, Gibson KJ, Pringle KG. Programming of Renal Development and Chronic Disease in Adult Life. Front Physiol. 2020;11:757.

33. Larkins N, Kim S, Craig J, Hodson E. Steroid-sensitive nephrotic syndrome: an evidence-based update of immunosuppressive treatment in children. Arch Dis Child. 2016;101(4):404-8.

34. Vivarelli M, Moscaritolo E, Tsalkidis A, Massella L, Emma F. Time for initial response to steroids is a major prognostic factor in idiopathic nephrotic syndrome. J Pediatr. 2010;156(6):965-71.

35. Colucci M, Carsetti R, Cascioli S, Casiraghi F, Perna A, Rava L, et al. B Cell Reconstitution after Rituximab Treatment in Idiopathic Nephrotic Syndrome. J Am Soc Nephrol. 2016;27(6):1811-22.

36. Tullus K, Webb H, Bagga A. Management of steroid-resistant nephrotic syndrome in children and adolescents. Lancet Child Adolesc Health. 2018;2(12):880-90.

37. Faul C, Donnelly M, Merscher-Gomez S, Chang YH, Franz S, Delfgaauw J, et al. The actin cytoskeleton of kidney podocytes is a direct target of the antiproteinuric effect of cyclosporine A. Nat Med. 2008;14(9):931-8.

38. Mekahli D, Liutkus A, Ranchin B, Yu A, Bessenay L, Girardin E, et al. Long-term outcome of idiopathic steroid-resistant nephrotic syndrome: a multicenter study. Pediatr Nephrol. 2009;24(8):1525-32.

39. Trautmann A, Schnaidt S, Lipska-Zietkiewicz BS, Bodria M, Ozaltin F, Emma F, et al. Long-Term Outcome of Steroid-Resistant Nephrotic Syndrome in Children. J Am Soc Nephrol. 2017;28(10):3055-65.

40. Cheong HI, Han HW, Park HW, Ha IS, Han KS, Lee HS, et al. Early recurrent nephrotic syndrome after renal transplantation in children with focal segmental glomerulosclerosis. Nephrol Dial Transplant. 2000;15(1):78-81.

41. Vivarelli M, Massella L, Ruggiero B, Emma F. Minimal Change Disease. Clin J Am Soc Nephrol. 2017;12(2):332-45.

42. Early identification of frequent relapsers among children with minimal change nephrotic syndrome. A report of the International Study of Kidney Disease in Children. J Pediatr. 1982;101(4):514-8.

43. Niaudet P. Long-term outcome of children with steroid-sensitive idiopathic nephrotic syndrome. Clin J Am Soc Nephrol. 2009;4(10):1547-8.

44. Saleem MA. Molecular stratification of idiopathic nephrotic syndrome. Nat Rev Nephrol. 2019;15(12):750-65.

45. Hamiwka LA, Midgley JP, Wade AW, Martz KL, Grisaru S. Outcomes of kidney transplantation in children with nephronophthisis: an analysis of the North American Pediatric Renal Trials and Collaborative Studies (NAPRTCS) Registry. Pediatr Transplant. 2008;12(8):878-82.

46. Gee HY, Ashraf S, Wan X, Vega-Warner V, Esteve-Rudd J, Lovric S, et al. Mutations in EMP2 cause childhood-onset nephrotic syndrome. Am J Hum Genet. 2014;94(6):884-90.

47. Dorval G, Gribouval O, Martinez-Barquero V, Machuca E, Tete MJ, Baudouin V, et al. Clinical and genetic heterogeneity in familial steroid-sensitive nephrotic syndrome. Pediatr Nephrol. 2018;33(3):473-83.

48. Ashraf S, Kudo H, Rao J, Kikuchi A, Widmeier E, Lawson JA, et al. Mutations in six nephrosis genes delineate a pathogenic pathway amenable to treatment. Nat Commun. 2018;9(1):1960.

49. Karp AM, Gbadegesin RA. Genetics of childhood steroid-sensitive nephrotic syndrome. Pediatr Nephrol. 2017;32(9):1481-8.

50. Robson KJ, Ooi JD, Holdsworth SR, Rossjohn J, Kitching AR. HLA and kidney disease: from associations to mechanisms. Nat Rev Nephrol. 2018;14(10):636-55.

51. Dendrou CA, Petersen J, Rossjohn J, Fugger L. HLA variation and disease. Nat Rev Immunol. 2018;18(5):325-39.

52. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, et al. The structure of haplotype blocks in the human genome. Science. 2002;296(5576):2225-9.

53. Kim AH, Chung JJ, Akilesh S, Koziell A, Jain S, Hodgin JB, et al. B cell-derived IL-4 acts on podocytes to induce proteinuria and foot process effacement. JCI Insight. 2017;2(21).

54. Gbadegesin RA, Adeyemo A, Webb NJ, Greenbaum LA, Abeyagunawardena A, Thalgahagoda S, et al. HLA-DQA1 and PLCG2 Are Candidate Risk Loci for Childhood-Onset Steroid-Sensitive Nephrotic Syndrome. J Am Soc Nephrol. 2015;26(7):1701-10.

55. Jia X, Horinouchi T, Hitomi Y, Shono A, Khor SS, Omae Y, et al. Strong Association of the HLA-DR/DQ Locus with Childhood Steroid-Sensitive Nephrotic Syndrome in the Japanese Population. J Am Soc Nephrol. 2018;29(8):2189-99.

56. Dufek S, Cheshire C, Levine AP, Trompeter RS, Issler N, Stubbs M, et al. Genetic Identification of Two Novel Loci Associated with Steroid-Sensitive Nephrotic Syndrome. J Am Soc Nephrol. 2019;30(8):1375-84.

57. Sen ES, Dean P, Yarram-Smith L, Bierzynska A, Woodward G, Buxton C, et al. Clinical genetic testing using a custom-designed steroid-resistant nephrotic syndrome gene panel: analysis and recommendations. J Med Genet. 2017;54(12):795-804.

58. Benoit G, Machuca E, Antignac C. Hereditary nephrotic syndrome: a systematic approach for genetic testing and a review of associated podocyte gene mutations. Pediatr Nephrol. 2010;25(9):1621-32.

59. Lowik MM, Groenen PJ, Levtchenko EN, Monnens LA, van den Heuvel LP. Molecular genetic analysis of podocyte genes in focal segmental glomerulosclerosis--a review. Eur J Pediatr. 2009;168(11):1291-304.

60. Bierzynska A, McCarthy HJ, Soderquest K, Sen ES, Colby E, Ding WY, et al. Genomic and clinical profiling of a national nephrotic syndrome cohort advocates a precision medicine approach to disease management. Kidney Int. 2017;91(4):937-47.

61. Dorval G, Kuzmuk V, Gribouval O, Welsh GI, Bierzynska A, Schmitt A, et al. TBC1D8B Loss-of-Function Mutations Lead to X-Linked Nephrotic Syndrome via Defective Trafficking Pathways. Am J Hum Genet. 2019;104(2):348-55.

62. Bensimhon AR, Williams AE, Gbadegesin RA. Treatment of steroid-resistant nephrotic syndrome in the genomic era. Pediatr Nephrol. 2019;34(11):2279-93.

63. Veissi S, Smeets B, van den Heuvel LP, Schreuder MF, Jansen J. Nephrotic syndrome in a dish: recent developments in modeling in vitro. Pediatr Nephrol. 2020;35(8):1363-72.

64. Gillies CE, Putler R, Menon R, Otto E, Yasutake K, Nair V, et al. An eQTL Landscape of Kidney Tissue in Human Nephrotic Syndrome. Am J Hum Genet. 2018;103(2):232-44.

65. Gessler M, Poustka A, Cavenee W, Neve RL, Orkin SH, Bruns GA. Homozygous deletion in Wilms tumours of a zinc-finger gene identified by chromosome jumping. Nature. 1990;343(6260):774-8.

66. Pelletier J, Bruening W, Kashtan CE, Mauer SM, Manivel JC, Striegel JE, et al. Germline mutations in the Wilms' tumor suppressor gene are associated with abnormal urogenital development in Denys-Drash syndrome. Cell. 1991;67(2):437-47.

67. Baird PN, Santos A, Groves N, Jadresic L, Cowell JK. Constitutional mutations in the WT1 gene in patients with Denys-Drash syndrome. Hum Mol Genet. 1992;1(5):301-5.

68. Frasier SD, Bashore RA, Mosier HD. Gonadoblastoma Associated with Pure Gonadal Dysgenesis in Monozygous Twins. J Pediatr. 1964;64:740-5.

69. Barbaux S, Niaudet P, Gubler MC, Grunfeld JP, Jaubert F, Kuttenn F, et al. Donor splice-site mutations in WT1 are responsible for Frasier syndrome. Nat Genet. 1997;17(4):467-70.

70. Denamur E, Bocquet N, Mougenot B, Da Silva F, Martinat L, Loirat C, et al. Mother-to-child transmitted WT1 splice-site mutation is responsible for distinct glomerular diseases. J Am Soc Nephrol. 1999;10(10):2219-23.

71. Koziell A, Grundy R. Frasier and Denys-Drash syndromes: different disorders or part of a spectrum? Arch Dis Child. 1999;81(4):365-9.

72. Kestila M, Lenkkeri U, Mannikko M, Lamerdin J, McCready P, Putaala H, et al. Positionally cloned gene for a novel glomerular protein--nephrin--is mutated in congenital nephrotic syndrome. Mol Cell. 1998;1(4):575-82.

73. Boute N, Gribouval O, Roselli S, Benessy F, Lee H, Fuchshuber A, et al. NPHS2, encoding the glomerular protein podocin, is mutated in autosomal recessive steroid-resistant nephrotic syndrome. Nat Genet. 2000;24(4):349-54.

74. Beltcheva O, Martin P, Lenkkeri U, Tryggvason K. Mutation spectrum in the nephrin gene (NPHS1) in congenital nephrotic syndrome. Hum Mutat. 2001;17(5):368-73.

75. Tory K, Menyhard DK, Woerner S, Nevo F, Gribouval O, Kerti A, et al. Mutation-dependent recessive inheritance of NPHS2-associated steroid-resistant nephrotic syndrome. Nat Genet. 2014;46(3):299-304.

76. Simons M, Schwarz K, Kriz W, Miettinen A, Reiser J, Mundel P, et al. Involvement of lipid rafts in nephrin phosphorylation and organization of the glomerular slit diaphragm. Am J Pathol. 2001;159(3):1069-77.

77. Zenker M, Aigner T, Wendler O, Tralau T, Muntefering H, Fenski R, et al. Human laminin beta2 deficiency causes congenital nephrosis with mesangial sclerosis and distinct eye abnormalities. Hum Mol Genet. 2004;13(21):2625-32.

78. Hasselbacher K, Wiggins RC, Matejas V, Hinkes BG, Mucha B, Hoskins BE, et al. Recessive missense mutations in LAMB2 expand the clinical spectrum of LAMB2-associated disorders. Kidney Int. 2006;70(6):1008-12.

79. Jarad G, Cunningham J, Shaw AS, Miner JH. Proteinuria precedes podocyte abnormalities inLamb2-/- mice, implicating the glomerular basement membrane as an albumin barrier. J Clin Invest. 2006;116(8):2272-9.

80. Ha TS. Genetics of hereditary nephrotic syndrome: a clinical review. Korean J Pediatr. 2017;60(3):55-63.

81. Hinkes BG, Mucha B, Vlangos CN, Gbadegesin R, Liu J, Hasselbacher K, et al. Nephrotic syndrome in the first year of life: two thirds of cases are caused by

mutations in 4 genes (NPHS1, NPHS2, WT1, and LAMB2). Pediatrics. 2007;119(4):e907-19.

82. Sadowski CE, Lovric S, Ashraf S, Pabst WL, Gee HY, Kohl S, et al. A single-gene cause in 29.5% of cases of steroid-resistant nephrotic syndrome. J Am Soc Nephrol. 2015;26(6):1279-89.

83. Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NS, et al. Human Gene Mutation Database (HGMD): 2003 update. Hum Mutat. 2003;21(6):577-81.

84. Genovese G, Friedman DJ, Ross MD, Lecordier L, Uzureau P, Freedman BI, et al. Association of trypanolytic ApoL1 variants with kidney disease in African Americans. Science. 2010;329(5993):841-5.

85. Genomes Project C, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, et al. A map of human genome variation from population-scale sequencing. Nature. 2010;467(7319):1061-73.

86. Friedman DJ, Pollak MR. Apolipoprotein L1 and Kidney Disease in African Americans. Trends Endocrinol Metab. 2016;27(4):204-15.

87. Kopp JB, Nelson GW, Sampath K, Johnson RC, Genovese G, An P, et al. APOL1 genetic variants in focal segmental glomerulosclerosis and HIV-associated nephropathy. J Am Soc Nephrol. 2011;22(11):2129-37.

88. Reiner AP, Susztak K. APOL1 Variants: From Parasites to Kidney Function to Cardiovascular Disease. Arterioscler Thromb Vasc Biol. 2016;36(2):219-20.

89. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409(6822):860-921.

90. Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, et al. Accurate multiplex polony sequencing of an evolved bacterial genome. Science. 2005;309(5741):1728-32.

91. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005;437(7057):376-80.

92. Rabbani B, Mahdieh N, Hosomichi K, Nakaoka H, Inoue I. Next-generation sequencing: impact of exome sequencing in characterizing Mendelian disorders. J Hum Genet. 2012;57(10):621-32.

93. Kiezun A, Garimella K, Do R, Stitziel NO, Neale BM, McLaren PJ, et al. Exome sequencing and the genetic basis of complex traits. Nat Genet. 2012;44(6):623-30.

94. Stenson PD, Mort M, Ball EV, Chapman M, Evans K, Azevedo L, et al. The Human Gene Mutation Database (HGMD((R))): optimizing its use in a clinical diagnostic or research setting. Hum Genet. 2020;139(10):1197-207.

95. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 2016;17(6):333-51.

96. Illumina. Explore illumina sequencing technology 2021 [Available from: https://www.illumina.com/science/technology/next-generation-sequencing/sequencing-technology.html.

97. van Dijk EL, Jaszczyszyn Y, Naquin D, Thermes C. The Third Revolution in Sequencing Technology. Trends Genet. 2018;34(9):666-81.

98. Hedges DJ, Burges D, Powell E, Almonte C, Huang J, Young S, et al. Exome sequencing of a multigenerational human pedigree. PLoS One. 2009;4(12):e8232.

99. Southan C. Has the yo-yo stopped? An assessment of human protein-coding gene number. Proteomics. 2004;4(6):1712-26.

100. Lee H, Deignan JL, Dorrani N, Strom SP, Kantarci S, Quintero-Rivera F, et al. Clinical exome sequencing for genetic identification of rare Mendelian disorders. JAMA. 2014;312(18):1880-7.

101. Lelieveld SH, Spielmann M, Mundlos S, Veltman JA, Gilissen C. Comparison of Exome and Genome Sequencing Technologies for the Complete Capture of Protein-Coding Regions. Hum Mutat. 2015;36(8):815-22.

102. Turro E, Astle WJ, Megy K, Graf S, Greene D, Shamardina O, et al. Whole-genome sequencing of patients with rare diseases in a national health system. Nature. 2020;583(7814):96-102.

103. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68-74.

104. Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, et al. Genetic structure of human populations. Science. 2002;298(5602):2381-5.

105. Campbell MC, Tishkoff SA. The evolution of human genetic and phenotypic variation in Africa. Curr Biol. 2010;20(4):R166-73.

106. Hess JF, Kohl TA, Kotrova M, Ronsch K, Paprotka T, Mohr V, et al. Library preparation for next generation sequencing: A review of automation strategies. Biotechnol Adv. 2020;41:107537.

107. Chilamakuri CS, Lorenz S, Madoui MA, Vodak D, Sun J, Hovig E, et al. Performance comparison of four exome capture systems for deep sequencing. BMC Genomics. 2014;15:449.

108. Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet. 2010;11(1):31-46.

109. Hardwick SA, Deveson IW, Mercer TR. Reference standards for next-generation sequencing. Nat Rev Genet. 2017;18(8):473-84.

110. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008;456(7218):53-9.

111. Deorowicz S, Grabowski S. Compression of DNA sequence reads in FASTQ format. Bioinformatics. 2011;27(6):860-2.

112. Cock PJ, Fields CJ, Goto N, Heuer ML, Rice PM. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res. 2010;38(6):1767-71.

113. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-9.

114. Stitziel NO, Kiezun A, Sunyaev S. Computational and statistical approaches to analyzing variants identified by exome sequencing. Genome Biol. 2011;12(9):227.

115. Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science. 2012;337(6090):64-9.

116. Chen J, Li X, Zhong H, Meng Y, Du H. Systematic comparison of germline variant calling pipelines cross multiple next-generation sequencers. Sci Rep. 2019;9(1):9345.

117. Chakravorty S, Hegde M. Gene and Variant Annotation for Mendelian Disorders in the Era of Advanced Sequencing Technologies. Annu Rev Genomics Hum Genet. 2017;18:229-56.

118. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461(7265):747-53.

119. MacArthur DG, Manolio TA, Dimmock DP, Rehm HL, Shendure J, Abecasis GR, et al. Guidelines for investigating causality of sequence variants in human disease. Nature. 2014;508(7497):469-76.

120. Lightbody G, Haberland V, Browne F, Taggart L, Zheng H, Parkes E, et al. Review of applications of high-throughput sequencing in personalized medicine: barriers and facilitators of future progress in research and clinical application. Brief Bioinform. 2019;20(5):1795-811.

121. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med. 2015;17(5):405-24.

122. Bean LJH, Funke B, Carlston CM, Gannon JL, Kantarci S, Krock BL, et al. Diagnostic gene sequencing panels: from design to report-a technical standard of the American College of Medical Genetics and Genomics (ACMG). Genet Med. 2020;22(3):453-61.

123. Simpson MA, Irving MD, Asilmaz E, Gray MJ, Dafou D, Elmslie FV, et al. Mutations in NOTCH2 cause Hajdu-Cheney syndrome, a disorder of severe and progressive bone loss. Nat Genet. 2011;43(4):303-5.

124. Power C, Elliott J. Cohort profile: 1958 British birth cohort (National Child Development Study). Int J Epidemiol. 2006;35(1):34-41.

125. Ma X, Shao Y, Tian L, Flasch DA, Mulder HL, Edmonson MN, et al. Analysis of error profiles in deep next-generation sequencing data. Genome Biol. 2019;20(1):50.

126. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9):1297-303.

127. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27(15):2156-8.

128. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43(5):491-8.

129. Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Van der Auwera GA, et al. Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv. 2018:201178.

130. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38(16):e164.

131. Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, et al. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature. 2013;493(7431):216-20.

132. Karczewski KJ, Weisburd B, Thomas B, Solomonson M, Ruderfer DM, Kavanagh D, et al. The ExAC browser: displaying reference data information from over 60 000 exomes. Nucleic Acids Res. 2017;45(D1):D840-D5.

133. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alfoldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581(7809):434-43.

134. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet. 2013;Chapter 7:Unit7 20.

135. Sim NL, Kumar P, Hu J, Henikoff S, Schneider G, Ng PC. SIFT web server: predicting effects of amino acid substitutions on proteins. Nucleic Acids Res. 2012;40(Web Server issue):W452-7.

136. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. 2019;47(D1):D886-D94.

137. Raczy C, Petrovski R, Saunders CT, Chorny I, Kruglyak S, Margulies EH, et al. Isaac: ultra-fast whole-genome secondary analysis on Illumina sequencing platforms. Bioinformatics. 2013;29(16):2041-3.

138. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4:7.

139. Staples J, Qiao D, Cho MH, Silverman EK, University of Washington Center for Mendelian G, Nickerson DA, et al. PRIMUS: rapid reconstruction of pedigrees from genome-wide estimates of identity by descent. Am J Hum Genet. 2014;95(5):553-64.

140. Li H. Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics. 2014;30(20):2843-51.

141. Guo Y, Ye F, Sheng Q, Clark T, Samuels DC. Three-stage quality control strategies for DNA re-sequencing data. Brief Bioinform. 2014;15(6):879-89.

142. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in genome-wide association studies. Bioinformatics. 2010;26(22):2867-73.

143. International HapMap C. The International HapMap Project. Nature. 2003;426(6968):789-96.

144. Dupont B. Nomenclature for factors of the HLA system, 1987. Decisions of the Nomenclature Committee on Leukocyte Antigens, which met in New York on November 21-23, 1987. Hum Immunol. 1989;26(1):3-14.

145. Marsh SG, Albert ED, Bodmer WF, Bontrop RE, Dupont B, Erlich HA, et al. Nomenclature for factors of the HLA system, 2010. Tissue Antigens. 2010;75(4):291-455.

146. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754-60.

147. MacDonald ME, Novelletto A, Lin C, Tagle D, Barnes G, Bates G, et al. The Huntington's disease candidate region exhibits many different haplotypes. Nat Genet. 1992;1(2):99-103.

148. Risch N. A note on multiple testing procedures in linkage analysis. Am J Hum Genet. 1991;48(6):1058-64.

149. Abecasis GR, Cherny SS, Cookson WO, Cardon LR. Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet. 2002;30(1):97-101.

150. Lander ES, Green P. Construction of multilocus genetic linkage maps in humans. Proc Natl Acad Sci U S A. 1987;84(8):2363-7.

151. Ott J, Wang J, Leal SM. Genetic linkage analysis in the age of whole-genome sequencing. Nat Rev Genet. 2015;16(5):275-84.

152. Kong A, Cox NJ. Allele-sharing models: LOD scores and accurate linkage tests. Am J Hum Genet. 1997;61(5):1179-88.

153. Asimit J, Zeggini E. Rare variant association analysis methods for complex traits. Annu Rev Genet. 2010;44:293-308.

154. Fuentes Fajardo KV, Adams D, Program NCS, Mason CE, Sincan M, Tifft C, et al. Detecting false-positive signals in exome sequencing. Hum Mutat. 2012;33(4):609-13.

155. Tam V, Patel N, Turcotte M, Bosse Y, Pare G, Meyre D. Benefits and limitations of genome-wide association studies. Nat Rev Genet. 2019;20(8):467-84.

156. Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet. 2014;10(5):e1004383.

157. Pe'er I, Yelensky R, Altshuler D, Daly MJ. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. Genet Epidemiol. 2008;32(4):381-5.

158. Almasy L. The role of phenotype in gene discovery in the whole genome sequencing era. Hum Genet. 2012;131(10):1533-40.

159. Pokrajac D, Kamber AH, Karasalihovic Z. Children with Steroid-Resistant Nephrotic Syndrome: A Single -Center Experience. Mater Sociomed. 2018;30(2):84-8.

160. Liu KA, Mager NA. Women's involvement in clinical trials: historical perspective and future implications. Pharm Pract (Granada). 2016;14(1):708.

161. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014;46(3):310-5.

162. Freedman BI, Kopp JB, Sampson MG, Susztak K. APOL1 at 10 years: progress and next steps. Kidney Int. 2021;99(6):1296-302.

163. Robinson JT, Thorvaldsdottir H, Wenger AM, Zehir A, Mesirov JP. Variant Review with the Integrative Genomics Viewer. Cancer Res. 2017;77(21):e31-e4.

164. Karle SM, Uetz B, Ronner V, Glaeser L, Hildebrandt F, Fuchshuber A. Novel mutations in NPHS2 detected in both familial and sporadic steroid-resistant nephrotic syndrome. J Am Soc Nephrol. 2002;13(2):388-93.

165. Caridi G, Bertelli R, Carrea A, Di Duca M, Catarsi P, Artero M, et al. Prevalence, genetics, and clinical features of patients carrying podocin mutations in steroid-resistant nonfamilial focal segmental glomerulosclerosis. J Am Soc Nephrol. 2001;12(12):2742-6.

166. Roselli S, Moutkine I, Gribouval O, Benmerah A, Antignac C. Plasma membrane targeting of podocin through the classical exocytic pathway: effect of NPHS2 mutations. Traffic. 2004;5(1):37-44.

167. Isojima T, Harita Y, Furuyama M, Sugawara N, Ishizuka K, Horita S, et al. LMX1B mutation with residual transcriptional activity as a cause of isolated glomerulopathy. Nephrol Dial Transplant. 2014;29(1):81-8.

168. Hall G, Lane B, Chryst-Ladd M, Wu G, Lin JJ, Qin X, et al. Dysregulation of WTI (-KTS) is Associated with the Kidney-Specific Effects of the LMX1B R246Q Mutation. Sci Rep. 2017;7:39933.

169. Braun DA, Lovric S, Schapiro D, Schneider R, Marquez J, Asif M, et al. Mutations in multiple components of the nuclear pore complex cause nephrotic syndrome. J Clin Invest. 2018;128(10):4313-28.

170. Rosti RO, Sotak BN, Bielas SL, Bhat G, Silhavy JL, Aslanger AD, et al. Homozygous mutation in NUP107 leads to microcephaly with steroid-resistant nephrotic condition similar to Galloway-Mowat syndrome. J Med Genet. 2017;54(6):399-403.

171. Collins AJ, Foley RN, Chavers B, Gilbertson D, Herzog C, Johansen K, et al. 'United States Renal Data System 2011 Annual Data Report: Atlas of chronic kidney disease & end-stage renal disease in the United States. Am J Kidney Dis. 2012;59(1 Suppl 1):A7, e1-420.

172. Friedman DJ, Kozlitina J, Genovese G, Jog P, Pollak MR. Population-based risk assessment of APOL1 on renal disease. J Am Soc Nephrol. 2011;22(11):2098-105.

173. Limou S, Nelson GW, Kopp JB, Winkler CA. APOL1 kidney risk alleles: population genetics and disease associations. Adv Chronic Kidney Dis. 2014;21(5):426-33.

174. Takahashi S, Hiromura K, Tsukida M, Ohishi Y, Hamatani H, Sakurai N, et al. Nephrotic syndrome caused by immune-mediated acquired LCAT deficiency. J Am Soc Nephrol. 2013;24(8):1305-12.

175. Caridi G, Gigante M, Ravani P, Trivelli A, Barbano G, Scolari F, et al. Clinical features and long-term outcome of nephrotic syndrome associated with heterozygous NPHS1 and NPHS2 mutations. Clin J Am Soc Nephrol. 2009;4(6):1065-72.

176. Ruf RG, Lichtenberger A, Karle SM, Haas JP, Anacleto FE, Schultheiss M, et al. Patients with mutations in NPHS2 (podocin) do not respond to standard steroid treatment of nephrotic syndrome. J Am Soc Nephrol. 2004;15(3):722-32.

177. Gale DP, Oygar DD, Lin F, Oygar PD, Khan N, Connor TM, et al. A novel COL4A1 frameshift mutation in familial kidney disease: the importance of the C-terminal NC1 domain of type IV collagen. Nephrol Dial Transplant. 2016;31(11):1908-14.

178. Plaisier E, Gribouval O, Alamowitch S, Mougenot B, Prost C, Verpont MC, et al. COL4A1 mutations and hereditary angiopathy, nephropathy, aneurysms, and muscle cramps. N Engl J Med. 2007;357(26):2687-95.

179. Wang M, Chun J, Genovese G, Knob AU, Benjamin A, Wilkins MS, et al. Contributions of Rare Gene Variants to Familial and Sporadic FSGS. J Am Soc Nephrol. 2019;30(9):1625-40.

180. Sauna ZE, Kimchi-Sarfaty C. Understanding the contribution of synonymous mutations to human disease. Nat Rev Genet. 2011;12(10):683-91.

181. Zhang F, Lupski JR. Non-coding genetic variants in human disease. Hum Mol Genet. 2015;24(R1):R102-10.

182. Do R, Kathiresan S, Abecasis GR. Exome sequencing and complex disease: practical aspects of rare variant association studies. Hum Mol Genet. 2012;21(R1):R1-9.

183. Barbitoff YA, Polev DE, Glotov AS, Serebryakova EA, Shcherbakova IV, Kiselev AM, et al. Systematic dissection of biases in whole-exome and whole-genome sequencing reveals major determinants of coding sequence coverage. Sci Rep. 2020;10(1):2057.

184. Yu H, Artomov M, Brahler S, Stander MC, Shamsan G, Sampson MG, et al. A role for genetic susceptibility in sporadic focal segmental glomerulosclerosis. J Clin Invest. 2016;126(4):1603.

185. Brown EJ, Schlondorff JS, Becker DJ, Tsukaguchi H, Tonna SJ, Uscinski AL, et al. Mutations in the formin gene INF2 cause focal segmental glomerulosclerosis. Nat Genet. 2010;42(1):72-6.

186. Boyer O, Nevo F, Plaisier E, Funalot B, Gribouval O, Benoit G, et al. INF2 mutations in Charcot-Marie-Tooth disease with glomerulopathy. N Engl J Med. 2011;365(25):2377-88.

187. Tey S, Shahrizaila N, Drew AP, Samulong S, Goh KJ, Battaloglu E, et al. Linkage analysis and whole exome sequencing reveals AHNAK2 as a novel genetic cause for autosomal recessive CMT in a Malaysian family. Neurogenetics. 2019;20(3):117-27.

188. Wang M, Li X, Zhang J, Yang Q, Chen W, Jin W, et al. AHNAK2 is a Novel Prognostic Marker and Oncogenic Protein for Clear Cell Renal Cell Carcinoma. Theranostics. 2017;7(5):1100-13.

189. Koziell A, Grech V, Hussain S, Lee G, Lenkkeri U, Tryggvason K, et al. Genotype/phenotype correlations of NPHS1 and NPHS2 mutations in nephrotic syndrome advocate a functional inter-relationship in glomerular filtration. Hum Mol Genet. 2002;11(4):379-88.

190. Kirby A, Gnirke A, Jaffe DB, Baresova V, Pochet N, Blumenstiel B, et al. Mutations causing medullary cystic kidney disease type 1 lie in a large VNTR in MUC1 missed by massively parallel sequencing. Nat Genet. 2013;45(3):299-303.

191. Han SW, Ahn JY, Lee S, Noh YS, Jung HC, Lee MH, et al. Gene expression network analysis of lymph node involvement in colon cancer identifies AHSA2, CDK10, and CWC22 as possible prognostic markers. Sci Rep. 2020;10(1):7170.

192. Kobayashi M, Chandrasekhar A, Cheng C, Martinez JA, Ng H, de la Hoz C, et al. Diabetic polyneuropathy, sensory neurons, nuclear structure and spliceosome alterations: a role for CWC22. Dis Model Mech. 2017;10(3):215-24.

193. Trimarchi H, Paulero M, Rengel T, Gonzalez-Hoyos I, Forrester M, Lombi F, et al. Mucin-1 Gene Mutation and the Kidney: The Link between Autosomal Dominant Tubulointerstitial Kidney Disease and Focal and Segmental Glomerulosclerosis. Case Rep Nephrol. 2018;2018:9514917.

194. Leroy X, Copin MC, Devisme L, Buisine MP, Aubert JP, Gosselin B, et al. Expression of human mucin genes in normal kidney and renal cell carcinoma. Histopathology. 2002;40(5):450-7.

195. Leroy X, Devisme L, Buisine MP, Copin MC, Aubert S, Gosselin B, et al. Expression of human mucin genes during normal and abnormal renal development. Am J Clin Pathol. 2003;120(4):544-50.

196. Gussow AB, Petrovski S, Wang Q, Allen AS, Goldstein DB. The intolerance to functional genetic variation of protein domains predicts the localization of pathogenic mutations within genes. Genome Biol. 2016;17:9.

197. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011;27(21):2987-93.

198. Merchant ML, Barati MT, Caster DJ, Hata JL, Hobeika L, Coventry S, et al. Proteomic Analysis Identifies Distinct Glomerular Extracellular Matrix in Collapsing Focal Segmental Glomerulosclerosis. J Am Soc Nephrol. 2020;31(8):1883-904.

199. Laskowski J, Renner B, Le Quintrec M, Panzer S, Hannan JP, Ljubanovic D, et al. Distinct roles for the complement regulators factor H and Crry in protection of the kidney from injury. Kidney Int. 2016;90(1):109-22.

200. Guo C, Liu Y, Zhao W, Wei S, Zhang X, Wang W, et al. Apelin promotes diabetic nephropathy by inducing podocyte dysfunction via inhibiting proteasome activities. J Cell Mol Med. 2015;19(9):2273-85.

201. Liu Y, Zhang J, Wang Y, Zeng X. Apelin involved in progression of diabetic nephropathy by inhibiting autophagy in podocytes. Cell Death Dis. 2017;8(8):e3006.

202. Yang FQ, Yang FP, Li W, Liu M, Wang GC, Che JP, et al. Foxl1 inhibits tumor invasion and predicts outcome in human renal cancer. Int J Clin Exp Pathol. 2014;7(1):110-22.

203. Miao Z, Balzer MS, Ma Z, Liu H, Wu J, Shrestha R, et al. Single cell regulatory landscape of the mouse kidney highlights cellular differentiation programs and disease targets. Nat Commun. 2021;12(1):2277.

204. Wuttke M, Li Y, Li M, Sieber KB, Feitosa MF, Gorski M, et al. A catalog of genetic loci associated with kidney function from analyses of a million individuals. Nat Genet. 2019;51(6):957-72.

205. Oz-Levi D, Olender T, Bar-Joseph I, Zhu Y, Marek-Yagel D, Barozzi I, et al. Noncoding deletions reveal a gene that is critical for intestinal function. Nature. 2019;571(7763):107-11.

206. Conomos MP, Miller MB, Thornton TA. Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. Genet Epidemiol. 2015;39(4):276-93.

207. Pal A, Kaskel F. History of Nephrotic Syndrome and Evolution of its Treatment. Front Pediatr. 2016;4:56.

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

209. Niemi MEK, Martin HC, Rice DL, Gallone G, Gordon S, Kelemen M, et al. Common genetic variants contribute to risk of rare severe neurodevelopmental disorders. Nature. 2018;562(7726):268-71.

210. Riise R, Andreasson S, Borgastrom MK, Wright AF, Tommerup N, Rosenberg T, et al. Intrafamilial variation of the phenotype in Bardet-Biedl syndrome. Br J Ophthalmol. 1997;81(5):378-85.

211. Hofmann C, Girschick H, Mornet E, Schneider D, Jakob F, Mentrup B. Unexpected high intrafamilial phenotypic variability observed in hypophosphatasia. Eur J Hum Genet. 2014;22(10):1160-4.

212. Harel T, Goldberg Y, Shalev SA, Chervinski I, Ofir R, Birk OS. Limb-girdle muscular dystrophy 2I: phenotypic variability within a large consanguineous Bedouin family associated with a novel FKRP mutation. Eur J Hum Genet. 2004;12(1):38-43.

213. Stokman MF, Renkema KY, Giles RH, Schaefer F, Knoers NV, van Eerde AM. The expanding phenotypic spectra of kidney diseases: insights from genetic studies. Nat Rev Nephrol. 2016;12(8):472-83.

214. Popejoy AB, Ritter DI, Crooks K, Currey E, Fullerton SM, Hindorff LA, et al. The clinical imperative for inclusivity: Race, ethnicity, and ancestry (REA) in genomics. Hum Mutat. 2018;39(11):1713-20.