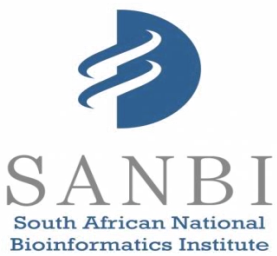
UNIVERSITY *of the*
WESTERN CAPE



Computational genomics approaches for kidney diseases in Africa

by

Darlington Shingirirai Mapiye

A thesis submitted in partial fulfilment for the degree of Doctor Philosophy at the South African National Bioinformatics Institute, Faculty of Science, University of the Western Cape

Supervisor: Dr. Nicki Tiffin

Co-Supervisor: Dr. Junaid Gamieldien

November 2015

Declaration

I, Darlington Shingirirai Mapiye, declare that this thesis titled, **"*Computational genomics approaches for kidney diseases in Africa*"** and the work presented in it is my own. I confirm that:

• This work was done wholly while in candidature for the degree of Doctor of Philosophy (PhD) at the University of the Western Cape and to the best of my knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the award of any other degree or diploma at University of the Western Cape or any other educational institution, except where due acknowledgement is made in the thesis.

• Any contribution made to the research by others, with whom I have worked with is explicitly acknowledged in the thesis

Signed:

Date:

Acknowledgments

Foremost, I would like to express my sincere gratitude to my advisor and mentor Dr Nicki Tiffin for the continuous support of my Ph.D. study and research, for her patience, excellent guidance, caring, motivation, enthusiasm, immense knowledge and providing me with an excellent atmosphere for doing research. Her guidance helped me in all the time of my research and writing of this thesis. I could not have imagined having a better advisor and mentor for my Ph.D. study.

Besides my advisor, I would like to thank Dr Galen Wright, Dr Reuben Cleote and Prof Ikechi Okpechi for their encouragement, insightful comments, involvement and hard questions. Special thanks go H3Abionet under the leadership of Professor Nicola Mulder for all the travel fellowships awarded to present my work at conferences, I couldn't have done it without their support. To my loving wife Marceline and son Dylan, thank you for your unwavering love and support. Finally, I would like to thank all the people who contributed in some way to the work described in this thesis not forgetting my family for their encouragement and love.

Abstract

End stage renal disease (ESRD), a more severe form of kidney disease, is considered to be a complex trait that may involve multiple processes which work together on a background of a significant genetic susceptibility. Black Africans have been shown to bear an unequal burden of this disease compared to white Europeans, Americans and Caucasians. Despite this, most of the genetic and epidemiological advances made in understanding the aetiology of kidney diseases have been done in other populations outside of sub-Saharan Africa (SSA). Very little research has been undertaken to investigate key genetic factors that drive ESRD in Africans compared to patients from rest of world populations.

Therefore, the primary aim of this Bioinformatics thesis was twofold: firstly, to develop and apply a whole exome sequencing (WES) analysis pipeline and use it to understand a genetic mechanism underlying ESRD in a South African population of mixed ancestry. As I hypothesized that the pipeline would enable the discovery of highly penetrate rare variants with large effect size, which are expected to explain an important fraction of the genetic aetiology and pathogenesis of ESRD in these African patients. Secondly, the aim was to develop and set up a multicenter clinical database that would capture a plethora of clinical data for patients with Lupus, one of the risk factors of ESRD.

From WES of six family members (five cases and one control); a total of 23 196 SNVs, 1445 insertions and 1340 deletions, overlapped amongst all affected family members. The variants were consistent with an autosomal dominant inheritance pattern inferred in this family. Of these, only 1550 SNVs, 67 insertions and 112 deletions were present in all affected family members but absent in the unaffected family member.

Following detailed evaluation of evidence for variant implication and pathogenicity, only 3 very rare heterozygous missense variants in 3 genes COL4A1 [p.R476W], ICAM1 [p.P352L], COL16A1 [p.T116M] were considered potentially disease causing. Computational relatedness analysis revealed approximate amount of DNA shared by family members and confirmed reported relatedness. Genotyping for the Y chromosome was additionally performed to assist in sample identity. The clinical database has been designed and is being piloted at Groote Schuur medical Hospital at the University of Cape Town. Currently, about 290 patients have already been entered in the registry.

The resources and methodologies developed in this thesis have the potential to contribute not only to the understanding of ESRD and its risk factors, but to the successful application of

WES in clinical practice. Importantly, it contributes significant information on the genetics of ESRD based on an African family and will also improve scientific infrastructure on the African continent. Clinical databasing will go a long way to enable clinicians to collect and store standardised clinical data for their patients.

Table of Contents

List of tables

List of figures

Publications

Publications arising from work in this thesis:

Bridget Hodkinson, **Darlington Mapiye**, David Jayne, Nicki Tiffin, Ikechi Okpechi. The African Lupus Genetics Network (ALUGEN) registry: standardized, prospective follow-up studies in African patients with Systemic Lupus Erythematous.

The candidate (Darlington S Mapiye) designed and implemented the clinical database and was involved in manuscript write-up.

Conferences and presentations

ISCB Africa and ASBCB conference. March 2015, Dar es Salaam, Tanzania

Oral Presentation:

Mapiye D, Galen Wright, Ikechi Okpechi. Computational genomic approaches for kidney diseases in Africa.

National Institutes of health (NIH) Consortium meeting. May 2015, Livingston, Zambia.

Poster Presentation:

Mapiye D, Galen Wright, Ikechi Okpechi. Computational genomic approaches for kidney diseases in Africa.

23rd International conference on intelligent systems molecular biology and14th European conference on molecular biology, July 2015, Dublin Ireland

Poster Presentation:

Mapiye D, Galen Wright, Ikechi Okpechi. Computational genomic approaches for kidney diseases in Africa.

Courses

Working with the human genome: Welcome Trust funded training workshop in Blantyre Malawi, January 2015.

Advanced Genome Wide Association (GWAS) modeling and simulation workshop. African Institutes of Mathematical Sciences, April 2015.

Medical population genetics and GWAS for complex diseases. African Institutes of Mathematical Sciences, April 2015.

Principles and Practice of Clinical Research. Medical Research Council and National Institutes of Health May 2015.

List of abbreviations

As	Alport syndrome
BAM	binary alignment map
BWA	burrow-wheeler aligner
CRF	case report forms
CWES	clinical whole exome sequencing
CGN	chronic glomerular nephritis
CKD	chronic kidney disease
CONIFER	copy number inference from exome reads
CNV	copy number variant
DNA	deoxy ribonucleic acid
HER	electronic health records
eGFR	estimated glomerular filtration rate
ESRD	end stage renal disease
FSGC	Focal segmental glomerular sclerosis
GBM	glomerular basement membrane
GATK	genome analysis tool kit
GERP	genomic evolutionary rate profiling
GWAS	genome wide association study
IVA	ingenuity variant analysis
IPA	ingenuity pathway analysis
INDEL	insertion deletion
JVC	joint variant calling
MALD	mapping by admixture linkage disequilibrium
MAF	minor allele frequency
NGS	next generation sequencing
POLYPHEN	polymorphism phenotyping
RRT	renal replacement therapy
REDCAP	research electronic data capture
SAM	sequence alignment map
SLE	systemic lupus erythematous
SIFT	sorting intolerant from to tolerant
SNP	single nucleotide polymorphism
SNV	single nucleotide variant
VEP	variant effect predictor
VNTR	variant number tandem repeat
WES	whole exome sequencing
VCF	variant call file

WGS	whole genome sequencing

1 Literature Review

1.1 Introduction

Amid rapid urbanisation, life style changes, and the increasing rates of non-communicable diseases, the sub-Saharan population is increasingly becoming vulnerable to chronic kidney diseases (CKD) (Stanifer et al., 2014). CKD affects an approximately 10 -13 % of adults in Sub-Saharan Africa (SSA), of these about 5-10 % reach end stage renal disease (ESRD), a more severe form of kidney disease which requires renal replacement therapy (RRT) to treat (Martins et al., 2012; Odubanjo et al., 2011; Schieppati and Remuzzi, 2005; Stanifer et al., 2014; Sumaili et al., 2009). Yet, only approximately 2% of the patients with ESRD are able to access this life saving treatment (RRT), making ESRD a death sentence for most patients (Abu-Aisha and Elamin, 2010; Katz et al., 2010).

ESRD is considered to be a complex trait that may involve multiple processes which work together on a background of a significant genetic susceptibility (Bowden, 2003). Black Africans have been shown to bear an unequal of this disease compared to white Europeans, Americans and Caucasians (Nugent et al., 2011). Despite this, most of the genetic and epidemiological advances made in the elucidation of the genetic aetiology of kidney diseases have been done in other populations outside of sub-Saharan Africa, mostly in African Americans and Europeans (Freedman et al., 1993; Genovese et al., 2010; Schieppati and Remuzzi, 2005). In addition, few epidemiological studies have been undertaken to ascertain the incidence, prevalence and other causes of CKD in developing countries. Thus, in order to address some of the affliction of CKD in Africa, the epidemiology of kidney disease needs to be established.

In collaboration with the Nephrology Unit at the Groote Schuur Teaching Hospital, Cape Town, South Africa. I have explored clinical applications of Bioinformatics tools, resources and research methodologies that can contribute to addressing the burden of ESRD in African populations. Two main applications were identified and explored further. The first problem I identified that can be addressed with Bioinformatics approaches is to understand the underlying disease mechanism,

especially in the case of unusual, idiopathic or extreme phenotypes. Very little research has been undertaken to investigate key genetic factors that drive ESRD in Africans compared to patients from rest of world populations. Omics approaches can be harnessed to better understand biological mechanisms that might be driving ESRD in African patients. In this study, I used whole exome sequencing to identify potentially causative variants for an unusual, difficult and severe autosomal dominant ESRD in a South African family of mixed ancestry, which is characterised by early onset elevated serum creatinine, and developmental defects, but with the absence of haematuria and proteinuria which are the commonly utilised clinical markers of renal insufficiency. Bioinformatics approaches were used to analyse and compare exome sequence data from six family members (5 affected and one unaffected).

The second clear problem identified is the poor unstructured collection, storage, accessibility and or reliability of clinical data collected from patients with kidney disease. In order to demonstrate effective application of clinical databasing which will result in collection and storage of reliable structured patient data, which can be used for future genomic studies, a multicentre clinical database has been developed.

Therefore, the primary aim of this Bioinformatics thesis was twofold: firstly, to develop and apply a whole exome sequencing (WES) analysis pipeline and use it to unravel and understand a genetic mechanism underlying ESRD in a South African population of mixed ancestry. As I hypothesised that the pipeline would enable the discovery of highly penetrate rare variants and other functional mutations with large effect size, which are expected to explain an important fraction of the genetic aetiology and pathogenesis of ESRD in these African patients; therefore, having a potential clinical interest. This would assist us to better understand the genetic mechanisms and disease pathogenesis of one form of ESRD based on an African population. Secondly, the aim is to develop and set up a multicentre registry that would capture a plethora of clinical data for patients with Lupus, one of the risk factors of ESRD. Lack of registries was identified as one of the major obstacles to obtaining reliable statistics about the prevalence and incidences of kidney diseases in many African countries, because registries offer an important source of information on multiple aspects of a disease. They are primarily useful in characterising disease population, describing the prevalence and incidences, trends in mortality and

investigating relationships among patient demographics, exposures, treatment modalities and morbidity.

1.2 Kidney Disease in Africa

The kidneys are vital excretory organs and central to fluid, electrolyte and acid-base homeostasis in humans (Figure 1.1). Damage of the kidneys has severe consequences for systemic functions, growth and survival. CKD is the presence of kidney damage, manifested by abnormal albumin excretion or deteriorated kidney function that lasts longer than three months as quantified by measured or estimated glomerular filtration rate (eGFR) (Kopple, 2001). Progressive renal disease usually leads to the common end point (ESRD), which is characterised by a shrunken, fibrotic kidney. CKD poses great challenges of a plethora of management modalities.

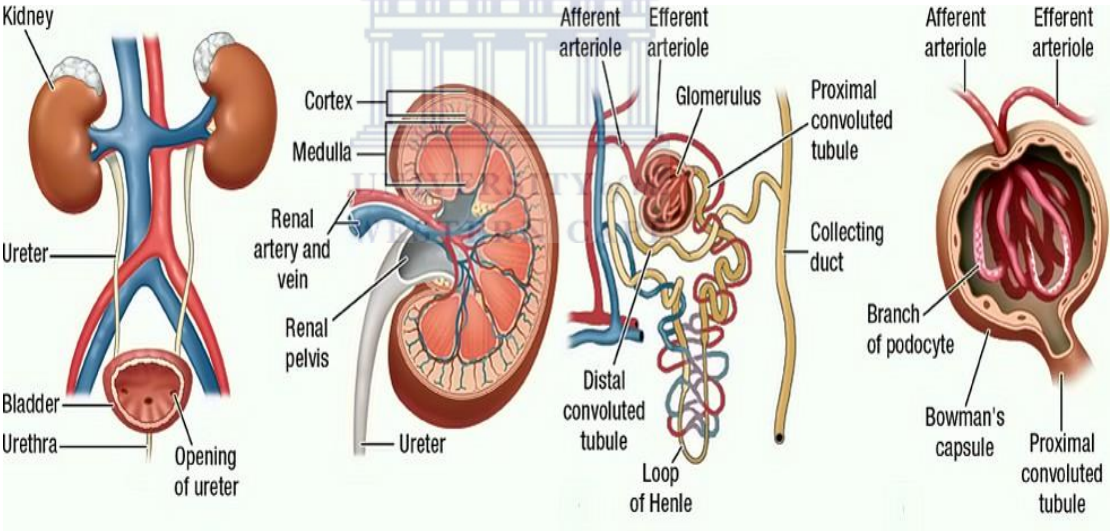


Figure 1.1: The human kidney and its functional components. (Ellsworth and Howard, 1934). A human being has two kidneys that are located in the lower abdomen. The kidneys receive most of their blood directly from the heart via the renal artery and the blood leaves the kidney via the renal vein. The main functional unit of the kidney is the nephron. Each kidney has over a million nephrons that contribute to its proper function. Each nephron consists of the glomerulus which is located in the cortex and passes its filtrate into the proximal convoluted tubules. The proximal tubule then leads to the loop of Henle, which is located in the medulla and it's mainly responsible for water reabsorption. This leads to the distal convoluted tubules that lead to the collecting duct. Blood for the nephron is supplied by afferent arteriole and leaves via the efferent arteriole.

1.2.1 Burden of kidney disease in African

Rapid urbanisation is occurring in many parts of Sub-Saharan Africa (SSA), contributing to over-crowding and poverty. While infectious and parasitic diseases are still the leading cause of death in SSA, non-communicable diseases are increasingly being recognised (Naicker, 2010). In 2011 the United Nations General Assembly accepted a resolution recognising the imminent risk of the non-communicable diseases and their affliction (Mensah and Mayosi, 2013). According to the Word Health report of 2002 and Global Burden of Diseases project, kidney disease contributes to the burden of diseases, with an approximately 850,000 deaths every year and over 15 million disability adjusted life years (Schieppati and Remuzzi, 2005).

CKD is now acknowledged as a global public health problem (Murray and Lopez, 1997; Stanifer et al., 2014). While the magnitude of CKD has been defined better in developed countries, growing evidence indicates that the burden of CKD is even greater in SSA than previously anticipated (Naicker, 2010). CKD and to a greater extent ESRD contribute substantially to the disparate burden of illness, disability and premature death across sex, age, race/ethnicity, socioeconomic status and geographic boundaries (Pugsley et al., 2009). Disadvantaged communities such as those in SSA, racial and ethnic minorities suffer from marked increases in incidences, prevalence and complications of CKD (Pugsley et al., 2009).

It is projected that by 2030 more than 70% of patients with ESRD will be living in low income countries, such as SSA, where the gross domestic product per person on average is less than 1,500 United States dollars per year (Barsoum, 2005). In SSA, public health care systems receive only 0.4 - 4% of the gross domestic product (GDP) and this has to be shared between infectious diseases such as the HIV aids pandemic and the emerging threat of non-communicable disease (Jafar et al., 2006). CKD is one of the serious health conditions that disproportionately afflict low income communities (such as those in SSA), their health systems as well as financial infrastructures (Jafar et al., 2006). However, much of the economic burden of CKD can be attributed to direct medical expenses associated with expensive long-treatment

costs (Pakistani, 1994).

A similar trend can be seen with ESRD management where developed countries dedicate more than 1% of the total health care budget to approximately 0.1% of population with ESRD, while in SSA ESRD management is too expensive and healthcare resources and budgets are unable to meet the burden of treatment (Hossain et al., 2009). For instance, in SSA, RRT costs are more than 10 times the annual per capita income and often health insurance coverage is low or non-existent for RRT treatment (Nugent et al., 2011). Thus, if the affected people do not receive RRT they will most likely die, creating another financial burden on the already resource limited countries, as their dependents would need to be taken care of, and their contributions to the economy are lost.

These exorbitant costs and lack of access to RRT are the major reason why approximately less than 10% of patients in low income countries receive RRT (Pakistani, 1994). Financial changes, reduced savings, decreased investment potential, constrained educational attainment that families face due to the burden of CKD result in detrimental socioeconomic impacts. These effects are likely to translate into significant decreased economic growth, and compounded over time would adversely affect disease management and control, in already resource poor settings (Organization and others, 2005).

Apart from the high costs of RRT, the pressure on national resources is further compounded by the high cardiovascular disease (CVD) burden observed in CKD patients. This is also exacerbated by the on-going brain drain of health workers, mainly physicians and nurses from Africa to more affluent regions. For instance, there are no nephrologists in many parts of SSA; the numbers vary from 0.5 per million populations (pmp) in Kenya to 0.7 pmp in Nigeria and 1.1 pmp in South Africa. This has a direct effect on availability of RRT as there will be no skilled personnel to oversee the therapy (Naicker, 2010).

The accumulative prevalence and incidence of CKD presents a worrying health burden in SSA. The surge in CKD and progression to ESRD mainly results from rising diabetes and hypertension pandemics (Murray and Lopez, 1997). This is

creating pressure on the already burdened health care system. Furthermore, limited access to health care, lack of awareness and limited capacity of health care workers suggest that those in lowest socioeconomic brackets are often oblivious to the risk of CKD and this adversely affects the outcome of the disease (Naicker, 2010). Treatment of ESRD is low priority for the already overwhelmed public health infrastructure (Naicker, 2010).

1.2.2 Epidemiological patterns of kidney disease in Africa compared to other populations

Kidney disease is an escalating global epidemic that disproportionally affects the economic, social, and health outcomes of resource-poor and low income countries such as those in Africa (Beaglehole and Yach, 2003). While significant advances have ensued in the management of CDK/ESRD patients worldwide with substantial improvements in outcomes and clinical state, survival is still very poor in SSA where GDP per capita is low and budgetary allotment on health is inadequate (Stanifer et al., 2014).

The pattern of disease morbidity and mortality throughout the world is fluctuating both in the developed and the developing countries (Beaglehole and Yach, 2003). During the 20th century, infectious diseases were the major cause of death and disability. In this century, however, an epidemiological transition has occurred resulting in non-communicable, non-infectious diseases becoming the leading cause of mortality and morbidity around the world (Yach et al., 2004). This variation is echoed in the type of diseases causing CKD, their presentation and progression. To date, the main cause of ESRD is diabetes resulting from the global pandemic of type 2 diabetes (Yach et al., 2004). Its rate of progression is extraordinary, and it is predicted that the number of patients with type 2 diabetes around the world will double in the next 25 years (Yach et al., 2004). Consequently, this will lead to a corresponding escalation in the number of patients with CKD and subsequently the number requiring RRT.

The epidemiology of kidney disease is strikingly different in SSA compared to developed countries (DuBose, 2007). While it predominately affects middle aged and elderly population in developed countries, in SSA it affects mostly young adults aged 20-50 years in their prime and most economically productive years (Mabayoje et al., 1992; Naicker, 1998). Other factors that may contribute include; poor access to health care, poor knowledge of the risk factors as well as detrimental socio-cultural practices (Arogundade et al., 2011). In contrast, the US prevalence of CKD escalates strongly with age (4% at age 29-39 y; 47% at age >70 y), with the most rapid growth in people aged 60 years or older (US, 1994). In the National Health and Nutrition Examination Survey (NHANES) study, the prevalence of stage 3 CKD in this age group rose from 18.8% during 1988−1994 to 24.5% from 2003-2006 (US, 1994). Throughout this same period, the prevalence of CKD in people aged 20-39 years remained consistently beneath 0.5 % and men and women showed similar prevalence (US, 1994).

On the other hand, the progression of CKD to the more severe ESRD has been reported to be rapid in Africa as compared to the USA and Europe. Within the USA, the prevalence of early CKD is comparable across racial/ethnic categories but the progression to ESRD is far more rapid among minority populations, with ESRD rates nearly 4 fold higher among African Americans in comparison to US white (US, 1994). This occurs despite both population races having similar prevalence rates of early CKD. Important differences also exist in the frequency of specific causes of CKD among different races. In the Chronic Kidney Disease in Children (CKiD) Study, for example, glomerular disease was much more common among non-white persons (Furth et al., 2006). However, the rapid progression in African Americans can be attributed to lower socioeconomic status, lesser access to health care, excess exposure to environmental toxins and other disease risk factors (Martins et al., 2012). In Mexico, it is estimated that the prevalence of CKD is as high as 15.8% among high-risk, poor populations, with similar demographics characteristics to Africans (Correa-Rotter and Gonzalez-Michaca, 2005).

The mortality rates associated with CKD are remarkable. After adjustment for age, gender, race, comorbidity, and prior hospitalizations, mortality in patients with CKD in 2009 was 56% greater than that in patients without CKD (Reyes-Bahamonde et al., 2014). For patients with ESRD the adjusted mortality rate is 76% greater. Mortality

rates are consistently higher for men than for women and for black persons than for white individuals and patients of other races (Agnes et al., 2012). The highest mortality rate is within the first 6 months of initiating dialysis. Mortality then tends to improve over the next 6 months, before increasing progressively over the next 4 years. The 5-year survival rate for a patient undergoing long-term dialysis in the United States is approximately 35%, and approximately 25% in patients with diabetes (Herzog et al., 2002).

In a study by Jaar, mortality risk was elevated in patients with ESRD and congestive heart failure who received peritoneal dialysis compared with those who received hemodialysis (Jaar et al., 2005). Their Median survival time was approximately 20 months in patients receiving peritoneal dialysis as compared to 36.7 months in patients receiving hemodialysis. Compared with non-dialysis patients and individuals without kidney disease, patients with ESRD on dialysis have significantly increased mortality.

A healthy person aged 60 years can expect to live for more than 20 years, whereas the life expectancy of a patient aged 60 years who is starting hemodialysis is closer to 4 years (Jaar et al., 2005). Among patients aged 65 years or older who have ESRD, mortality rates are 6 times higher than in the general population (Rao et al., 2007). Puberty is often delayed among males and females with significant CKD (Seikaly et al., 2006). Female patients with advanced CKD commonly develop menstrual irregularities. Women with ESRD are typically amenorrheic and infertile (Anantharaman and Schmidt, 2007). However, pregnancy can occur and can be associated with accelerated renal decline, including in women with a kidney transplant in advanced CKD (Watnick, 2007).

1.2.3 Risk factors of kidney disease in Africa

Diabetes Mellitus has emerged as the major risk factor for CKD and the commonest cause of ESRD in developed countries, while chronic glomerulonephritis (CGN) and hypertension (HTN) are the major risk factors in SSA, reflecting the high prevalence of bacterial, viral and parasitic infections affecting the kidneys in Africa (Abboud et

al., 1989; Akinsola et al., 2004; Bamgboye, 2005; Matekole et al., 1993). However, with the prevalence of diabetes in developing countries(South Africa 14-20%, Egypt 13%, Sudan 9%) rapidly approaching that of developed countries and an estimated 366 million adults expected to have diabetes by year 2030, this presents as major risk for SSA (Martins et al., 2012). Several reports in Nigeria and other SSA countries have established that HTN and CGN are the leading causes of ESRD, but the prevalence of diabetic nephropathy is rising and toxic nephropathy also contribute significantly (Naicker, 2003). In South Africa, HTN affects approximately 25% of the adult population and is the leading cause of CKD in 21% of patience on RRT registry, and it was the major cause of ESRD in Black South Africans accounting for approximately 35% of the ESRD racial group (Veriava et al., 1990). In contrast, hypertension was reported to be the cause of ESRD in approximately 4% of the white South Africans, 20% Indians, clearly showing the risk that HTN poses on Black South Africans (Naicker, 2003).

In contrast, type 2 diabetes has been shown to be the commonest cause of ESRD globally, accounting for up to 40% of new cases of the disease (Beulens et al., 2010). This is corresponding with the global increase in the prevalence of obesity. The prevalence of obesity and diabetes in South Africa has been described to be high and that mortality from diabetes is expected to increase by 38% in the period from 1995 to 2006 with an even greater growth of 67% reported for mortality due to kidney diseases (Amos et al., 1997). Diabetic patients have been under-represented in registry data hence accurate data on the prevalence of diabetes in the South African ESRD population are lacking. A study in the Western Cape province indicated that less than 20% of diabetic patients evaluated for RRT between 1988 and 2003 actually accessed RRT, consequently diabetic patients only comprised 6.2% of accepted patients overall (Veriava et al., 1990).

CGN disease is also common in SSA and is a significant cause of ESRD. However, studies from different parts of SSA display differences in the prevalence patterns of glomerular injury. For example, in Nigeria children with the nephrotic syndrome, membraneproliferative patterns on biopsy dominate whereas in South Africa FSGS appears to be the commonest. Thus, glomerular disease in Africa is more prevalent and seems to be a more severe form than that found in developed countries and is

characterised by poor response to treatment and rapid progression to renal failure (ESRD) (Naicker, 2003).

HIV is another emerging risk factor to kidney disease in Africa. Reported prevalence of kidney disease in HIV infected patients in SSA ranges from approximately 6% to just below 50%. Screening studies in South Africa reported HIV associated nephropathy in 55 -83% on biopsy (Naicker, 2003).

Environmental pollution, pesticides, analgesic abuse, herbal medicines and unregulated food additives also contribute to the disproportionate burden of CKD in many African countries. This is also complemented by poor health infrastructure, lack of access to health care for those leaving in remote areas and the continuous brain drain of the much-needed medical personnel (Nugent et al., 2011).

1.3 Diagnosis and Management of kidney disease

According to the National Kidney Foundation Kidney Disease Outcomes Quality Initiative (KDOQI) guidelines, CKD is defined by the presence of renal damage or decreased function that persists longer than three months. The diagnosis of CKD may be undertaken by the use of blood or urine laboratory markers of kidney damage or abnormal renal function, or by demonstrating structural damage on imaging studies, or by pathologic change on renal biopsy. This includes: abnormalities of urinary sediment: red blood cell casts (glomerular injury), white cell casts (interstitial/tubular injury), unusual rate of albumin excretion (albuminuria) and/or reduced GFR, or radiographic imaging abnormalities: Change in size or contour of the kidneys, hydronephrosis, polycystic disease, papillary necrosis, and pathologic abnormalities on renal biopsy: Vascular disease, glomerulitis, tubulointerstitial damage (Kopple, 2001).

The simplest, most reliable and recognized technique used to identify renal damage is by testing for albuminuria. Excessive albumin excretion is a reflection of primary kidney disease or renal involvement by a systemic vascular disorder which may follow underlying diseases such as hypertension, diabetes, and atherosclerosis. In a

few patients, screening could be started by urinalysis dipstick testing for proteinuria, which if positive would need to be confirmed by some measure of the albumin excretion rate. Urinary albumin dipstick testing and the measurement of the urinary albumin-to-creatinine ratio can also be used to assess adults with CKD (Kopple, 2001).

Another alternative way to diagnose, detect and monitor abnormal kidney function is to measure or estimate GFR (Kopple, 2001). A determination of GFR ought to be done in all patients with renal disease or signs of impaired renal function. The GFR indicates extend of renal functional impairment, is a valuable guide to dosage adjustment of drugs cleared by the kidney, and can be used to follow the course of kidney disease and to assess the response to therapy. A GFR less than 60 mL/min/1.73 m^2 is diagnostic of CKD.

Although the presence of CKD can be established on the basis of albuminuria and reduced GFR, proper diagnosis also includes identifying the underlying cause, as this may have important therapeutic and other management implications. A host of etiologies can be responsible for renal damage and diminished function, including hypertension, diabetes, autoimmune diseases, glomerulonephritis, drug-induced nephritis, and lower urinary tract obstructive disorders (Kopple, 2001). Therefore, alternative histological diagnosis methods such as kidney biopsy need to be sought and used.

1.3.1 Use of kidney biopsy

Lack of diagnosis consistency and reliability rendered on the basis of clinical features alone is making diagnosis of kidney disease problematic and perilous, underlining the need for kidney biopsy (Bihl et al., 2006; Gladman et al., 1989; Nossent et al., 1991). Kidney biopsy can be of paramount importance to those patients where disease classification is based on histological diagnosis and disease progression can be mitigated by treatment.

In a study by Haider et al, it was established that early biopsy-guided commencement

of therapy had profound benefits for patients as it preserved kidney function (Haider et al., 2012). Importantly, it was also recognized that kidney biopsy should be considered for patients independent of their age (Haider et al., 2012). Renal biopsy has also become an important means of diagnosing, prognosticating and guiding treatment for CGN, one of the major causes of ESRD in Africa (Okpechi et al., 2010). The patterns of glomerular disease in America, Europe, and Asia are well known and published as compared to Africa, highlighting a lack of clinical data registries where such information might be stored (Covic et al., 2006; Okpechi et al., 2010; Swaminathan et al., 2006).

FSGC is the most common cause of nephrotic syndrome in black patients, and IgA nephropathy, a common glomerular disease worldwide, may mimic lupus nephritis with disparate prognosis and management; neither of the diseases can be diagnosed on clinical grounds (Maisonneuve et al., 2000). Therefore, tissue diagnosis by kidney biopsy takes on even greater importance when considering these diagnoses. In order to attain such diagnostic success, it is clear that a kidney biopsy is essential in establishing not only the diagnosis, but the prognosis, as well treatment guidance (Dhaun et al., 2014).

However, little is known about the patterns of renal disease in African countries, mainly due to nonexistent of renal biopsy registries or because renal biopsy as a tool for diagnosing renal disease is entirely unavailable (Okpechi et al., 2010). Thus, with the diverse renal histopathological findings possible in kidney disease patients, biopsy determines not only the diagnosis and prognosis, but also substantially guides the management of this complex disease. As the therapeutic armamentarium for kidney disease expands, it becomes even more imperative that the correct diagnosis be made prior to beginning therapy (Dhaun et al., 2014).

In deciding whether to perform a biopsy, one must balance the risks of the biopsy procedure against the risks of limited diagnostic information, which may result in progression of potentially preventable renal disease or the unnecessary use of a possibly toxic therapy (Dhaun et al., 2014). Any consideration of the benefits of kidney biopsy must include knowledge of the risks of the procedure. With improved imaging and the use of semi-automated biopsy guns, complications are uncommon;

however, bleeding remains a foremost concern (Bihl et al., 2006; Dhaun et al., 2014). Major complications, those requiring blood transfusion or invasive intervention, have been reported in 0–6.4% of biopsies (Bihl et al., 2006; Dhaun et al., 2014). Predictors of complications have included low hematocrit and high creatinine. Thus, the use of renal biopsy and establishment of clinical registry is becoming increasingly imperative.

1.3.2 Kidney disease management

The management of CKD is manifold, encompassing a series of strategic measures designed to reduce the risk of further damage and slow the progression of kidney disease. Detecting and treating reversible causes should be considered in any patient with unfamiliar etiology of kidney disease. For instance, optimal control of glucose in diabetic patients, blood pressure control in those that are hypertensive and initiation of ACE-I or ARB therapy, are key to reducing disease progression (Stevens and Levin, 2013). In diabetic patients or those receiving loop diuretics, nephrotoxic agents ought to be avoided at all costs. These include NSAIDs, aminoglycoside antibiotics, and radiographic contrast material. Other supplementary measures to protect the kidney and slow progression to ESRD include smoking cessation, statin therapy to control hyperlipidemia, dietary protein restriction, and satisfactory treatment of metabolic acidosis (Stevens and Levin, 2013).

For cases of CKD that do progress to ESRD, it is imperative to anticipate and prepare patients for RRT (Kopple, 2001). For instance, patients with a GFR less than 30 mL/min/1.73 m^2 should prepare for imminent ESRD and eventually RRT (Kopple, 2001). It is also essential to bear in mind that acute deteriorations in GFR are frequently due to reversible factors such as volume depletion, radiographic contrast or nephrotoxic drug use, and urinary tract obstruction. Efforts should therefore be undertaken to rectify these in order to properly address declines in GFR and ascertain if true progression of the disease has occurred (Kopple, 2001).

In order to deal with the common risk factors of kidney diseases a rigorous multifactorial management approach is vital. The mainstays of treatment are

management of complications and/or comorbidities, lifestyle modification, and dialysis for patients with severe or ESRD. Some patients may be candidates for kidney transplant, although the wait for a non-related donor can be long. Psychosocial issues and patient education, primarily to ensure compliance with the established treatment plan is important.

Given that so many different factors can contribute to so many different forms of CKD and ESRD, it is crucial to assemble as much information as possible about each case that presents in the clinic. Understanding the factors that contribute to effective diagnosis, disease aetiology, patient prognosis and therapeutic options is essential in order to ensure the best possible patient outcomes.

1.4 Clinical databasing

Despite the magnitude of problems caused by kidney disease in Sub-Saharan Africa (SSA), there is insufficient systematically collected clinical data on disease characteristics and long-term outcomes in patients with CKD (Jha et al., 2013). The lack of such systematically collected data presents a gap that needs to be urgently bridged as a crucial initial step towards confronting the burden of CKD and its risk factors, specifically in developing countries (Singh et al., 2012). Reliable data that can be drawn from these clinical registries might assist policy makers in low income countries to formulate strategies that can be used to improve diagnosis, treatment and management of kidney diseases, which may eventually lead to improved patient outcome (Okpechi et al., 2010).

A clinical database is any systematic compilation of data for the purpose of health care planning, implementation and evaluation in a well-defined population. The data compiled are periodically published as statistical information to describe and analyse the state of the health of the population. For instance, an analysis of patterns of renal disease in South Africa based on a renal biopsy database provided further proof of HIV renal disease in SSA and a motive for prevention, early detection and aggressive treatment (Okpechi et al., 2010). Also, based on the analysis of data from renal biopsy registries, it was established that kidney biopsy at stages 1 or 2 and consecutive

therapy preserves kidney function and prevents disease progression through early initiation of treatment (Haider et al., 2012). An analysis of data from several renal registries in the Asia–Pacific region illustrated the wide application of registry data for planning dialysis services, for evaluating dialysis practices and health outcomes, with a view to improving the quality of dialysis care (LIM et al., 2008). This evidently highlights the important advances that can be drawn from well set up clinical databases.

Clinical databases come in a variety of forms, differing by their target entities, population coverage, type of data collected and their principal uses. There are two main general types; firstly, patient registers which are organized systems that use observational study methods to collect uniform data to evaluate specified outcomes for a population defined by a particular disease or therapy (Gliklich and Dreyer, 2010). Secondly, disease registers which are continuous, systematic collections of data on all cases of a disease occurring in a defined population with the purpose of assessing and controlling the impact of the disease in the community (Porta et al., 2014). Disease registers are closely related to public health or disease surveillance (Porta et al., 2014). Disease registers compile individual case level data while disease surveillances obtain data on the target disease from a variety of sources in addition to individual cases (Porta et al., 2014).

Clinical databases can be used to perform clinical research on disease presentation, prognosis, and treatment effectiveness to contrast with treatment efficacy in clinical trial (Singh et al., 2012). Also, epidemiology research such as studies on disease occurrence and distribution, disease risk or etiology and disease prevention can be done using data from clinical databases. Furthermore, health economic research to evaluate the cost effectiveness of health-care intervention can also be done using clinical registries data (Singh et al., 2012).

Considering that one of the major risk factor of ESRD in SSA is CGN, a disease which requires histopathological diagnosis in order to be properly treated and managed (Arogundade et al., 2011) , it has become increasing imperative that a formal structured way of storing clinical data for patience with diseases such as CGN be sought. The application of clinical databasing as an alternative to alleviate or remedy

the situation then becomes vital. It has been noted that reliable statistics required to elucidate epidemiological patterns of kidney disease in SSA are difficult to obtain. Therefore, setting up of multicenter registries can go a long way in bridging the gap and provide such much needed data. Once data is available in a formally structured and secured datatabase then it becomes easier to analyze this data and provide valuable insights that may be used to inform allocation of resources, for example health workers and to also see which treatment regimens are working and for which patients. In other words, the establishment of these clinical registries is an area of clinical informatics research ought to be given some attention and this thesis addresses a part of this problem.

1.5 Genetics of Kidney disease

Throughout the past decades breakthroughs in molecular biology and genetics have set the stage for a revolution in medicine. Advances in gene cloning, gene mapping and mutation analysis have contributed to a massive explosion of new information regarding the fundamental biological and pathophysiological basis for hundreds of human diseases (Gonzalez-Angulo et al., 2010). Accompanying this wave of new evidence is the realisation that most human diseases are significantly impacted by genetic factors (Bowden, 2003). There is an amassed understanding of the impact of genetic variability on the development of CKD, which is becoming clearer and highlights the need to elucidate the genetic basis of renal disease and its complications (Bowden, 2003). This would enhance our understanding of the diverse phenotypes observed in kidney diseases and enable us to determine the genetic predisposition to terminal complications (Agrawal et al., 2010).

In nephrology, a comprehensive range of clinical phenotypes can now be explained at a molecular level. The greatest strides have been made in defining genes responsible for a variety of inherited kidney diseases, including polcystic kidney disease, Alport syndrome and Bartter syndrome (Al-Bhalal and Akhtar, 2005; Hudson, 2004). Several obstacles still hinder the development of reliable, clinically useful molecular diagnostic assays for inherited kidney diseases. Two such issues which combine to make molecular diagnostic approaches highly challenging are genetic heterogeneity

and allelic heterogeneity. Understanding key genetic changes underlying the disease phenotype can lead to a broader understanding of the physiological mechanisms that cause the disease. In this way identifying mutations in individual patients and families can lead to a better understanding of disease mechanisms in general. However, knowing the precise molecular genetic basis for a disease in an individual patient has great clinical utility. Such information can be useful for diagnostic and prognostic evaluations and may soon be important for directing specific therapy (Bowden, 2003). Currently, clinically applicable molecular genetic tests are only available for a small fraction of diseases and more needs to be done so as to unleash the power of genetics to improve the lives of patients with kidney disease. To date most of these advances have been made possible by our understanding of the human genome (Consortium and others, 2011).

1.5.1 Human genome

The human genome is composed of approximately 3 billion nucleotide base pairs arranged into nearly 30,000 genes. Each gene contains both protein-coding and non-coding regions. Coding regions (exons) contain information for the construction of the amino-acid sequence of the protein product and structural or regulatory RNA species. Non-coding regions include introns and the 3'- and 5' regions of each gene and are generally not translated. Most variation in humans happens in the non-coding DNA regions and in degenerate positions in amino acid codons that do not change the envisioned identity of the corresponding amino acid. Humans differ on average 1 out of 100 nucleotides and most of these disparities occur often in the region with slight or no effect on protein function. As such, they are called polymorphisms. Mutations in the genetic sequence are more likely to have damaging effects if they result in a shift of the reading frame of the protein coding sequence, non-synonymous substitution of one amino acid for another (particularly amino acids with vastly different chemical properties), insertion of a premature stop codon resulting in a truncation of the protein product, or loss of a stop codon leading to an inappropriately extended protein product. Though protein-coding genes occupy approximately 1% of the genome, this region holds almost 85% of currently identified mutations with large effects on disease-related traits (Consortium and others, 2012).

1.5.2 Genes and disease association

Connecting phenotype with genotype is the fundamental goal of genetics. Determining DNA sequence causes human disease remains difficult for genetics (Consortium, 2012). For most of the modern era of human genetics before the advent of whole genome and exome sequencing the principal method for the identification of disease-associated genes was linkage analysis (positional cloning) (Botstein et al., 1980). This method is not reliant on any prior knowledge of biology or function, and is instead based purely on the inheritance of a trait in combination with the inheritance of chromosomal regions to identify the location of disease-related genes (Botstein et al., 1980). Using this method one or more pedigrees in which the trait of interest is observed to segregate are used. DNA from both affected and unaffected individuals are genotyped for polymorphic markers spread throughout the genome. Making use of the recombination that occurs in meiosis, one can identify a chromosomal region that shows segregation of a disease associated haplotype in affected individuals and non-disease associated haplotype in unaffected individuals (Lathrop et al., 1985). The method typically identifies a genomic interval spanning 0.5-10 cM which could contain up to 300 genes. Classic examples of early successes of positional cloning in identifying disease-causing genes consist of hemochromatosis disease *(MHC)* and nail patella syndrome *(LMX1B)* (Dreyer et al., 1998; Feder et al., 1996)). Other successes included identifying genes underlying cystic fibrosis fanconi anemie (FACC) (Strathdee et al., 1992). In addition, genetics factors underlying pre-dispositioning to cancer such as retinoblastoma, breast cancer and polyposis colorectal cancer were also identified using this method (Nishisho et al., 1991; Wooster et al., 1995). Also, the gene for Huntington disease *(HTT)* was mapped using positional cloning (Andrew et al., 1993).

In consanguineous families with suspected autosomal recessive traits a form of linkage analysis (homozygosity mapping) is used (Theis et al., 2011). This strategy can identify genomic regions in which candidate genes can be tested for the presence of pathogenic mutations. Homozygosity mapping has recently been used in combination with high-density mapping whole genome genotyping to identify disease

genes in patients in whom homozygosity by descent is suspected (Molho-Pessach et al., 2012).

It is worth mentioning that although linkage analysis to ascertain the causal genes for Mendelian disorders is popular, other approaches are also possible. For example, the gene responsible for haemophilia A was determined based on rescue of the clotting function of blood by a "globulin" isolated from normal blood (Ingram, 1976). Also, a candidate gene list can be determined based on function instead of location (as with linkage analysis), and cases and controls sequenced directly for potential mutations. While many rare disorders are highly amenable to linkage analysis, however, some disorders present a challenge for these methods. First, those which are extremely rare have only a few affected individuals and families per disorder, which result in underpowered analyses and/or large regions under the linkage peak(s). Second, these disorders are rare because the causal mutations are of large effect and under strong negative selection. Therefore, these mutations are not often transmitted through many generations and are, in fact, likely to be new events. Since linkage analysis is completely inheritance-dependent, such events may not be ascertained at all (Brunham and Hayden, 2013).

The biological and medical significance of these disease gene discoveries cannot be overstated. Once a disease gene has been identified, a massive volume of information about the biological function of that gene is provided by the phenotype of individuals in whom it is dysfunctional. Conversely, study of biological pathways that the gene product is involved in illuminates disease pathophysiology that can be informative for closely related disease phenotypes. From a clinical perspective, identification of a disease gene opens the door to diagnostics and predictive testing where appropriate (Tibben, 2007). The task of identifying genes associated with disease will henceforth rapidly accelerate as it is now tremendously facilitated by the sequencing of the human genome and the advert of next generation sequencing techniques (Lander, 2011; Venter et al., 2001).

1.5.3 Evidence for a genetic component to kidney disease

Knowledge of the primary cause of a disease is crucial for understanding its mechanisms and for adequate classification, prognosis, and treatment. Multiple lines of evidence suggest that the aetiologies and susceptibly to develop kidney disease has a significant genetic component (Bowden, 2003; George Jr and Neilson, 2000). These studies include familial aggregation studies, comparisons of incidence rates between different racial or ethnic populations, segregation analysis and advanced genome wide analysis studies. There is increasing understanding of the impact of genetic variability on the development of renal failure, which is becoming clearer and emphasises the need to elucidate the genetic basis for kidney diseases and its complications (Brunham and Hayden, 2013). This may perhaps lead to better understanding of the different phenotypes observed in renal diseases such as ESRD and would enable us to determine the genetic pre dispositioning to terminal complications.

Using Mapping by Admixture linkage disequilibrium (MALD) a strong association between genetic variants in genes (MYH9 and APOL1) and kidney disease due to HTN and adult onset FSGS have been identified in African Americans (Freedman et al., 2010; Genovese et al., 2010; Iyengar et al., 2007; Kao et al., 2008). Furthermore, 13 loci and nearly 20 variants have been linked with kidney disease, FSG and nephrotic syndrome in children (Boyer et al., 2011; Hildebrandt et al., 1997; Kottgen et al., 2008). A meta-analysis of GWAS data from four population-based cohorts in which 2400 people had CKD identified a highly significant association with variants within the UMOD gene (Kottgen et al., 2008; Okada et al., 2012). These results were replicated in an independent population.

In the Framingham heart study, 16 candidate genes were identified (Kottgen et al., 2008). Further analysis found a gene methenyltetrahydrofolate synthesis (MTHFS) to be significantly associated with CKD, indicating the possible involvement of this gene in CKD. These results were replicated in approximately 15,000 patients of the Atherosclerosis in communities study (ARIC) (Kottgen et al., 2008). In another study, evidence of the association of KLB1 gene was observed in African Americans families with multiple cases of ESRD (Yu et al., 2000). Further analysis in this gene

identified 12 allelic variants; of interest was C699A polymorphism in the coding sequence which was observed in 8 families but not found in the control samples (Yu et al., 2000). A genome-wide search for linkage to CKD was performed in 848 Mexican Americans from 26 families. The results showed a linkage on chromosomal regions 2p25 and 9q21 to several parameters of renal function (creatinine clearance and eGFR) (Arar et al., 2008). The long arm of chromosome 7 was found to have strongest evidence of linkage in a study of 98 siblings with T2DM and renal failure. On the other hand, in a genome wide analysis study on 18 Turkish families with CKD, a strong linkage peak was observed on chromosome 18 (LOD score 6.6) (Sale and Freedman, 2006). Evaluation of this locus in Pima Indians showed evidence of confirmation. In 2006, a review conducted on 56 genes related to CKD identified 15 genes which are related to immunity and defence, two of the key mechanisms involved in CKD (Imperatore et al., 1998).

In studies of African Americans, family history of ESRD was determined, with the conclusion that the presence of a close relative with ESRD gave an African American an eight fold increased risk of developing ESRD (Freedman et al., 2010). In Caucasians the increased risk was 2.7-fold. Seaquist et al. published the first description of familial aggregation in renal disease, primarily studying Caucasian type1 diabetes (TIDM) families, and came to a conclusion that 83% of diabetic siblings of probands receiving kidney transplant had developed kidney disease (Seaquist et al., 1989). Similar results were obtained by Borch-Johnsen et al. in European study of TIDM families and American TIDM families. Familial clustering of kidney disease was also observed in the Diabetes control and complications trial (Borch-Johnsen et al., 1992). Kidney disease clustering in families has also been described in affected Pima Indians, African Americans and Causians with similar conclusions, that if a family member has renal disease then there is a higher risk of other family members getting the disease (Freedman et al., 2010; O'Dea et al., 1998; Pettitt et al., 1990). As it can be clearly seen, most of the genetics studies carried out to ascertain the genetic basis of CKD and ESRD are conducted in populations other than Africans, underlining the need to elucidate the genetic component of CKD based on an Africa population.

What makes it more interesting and potentially more informative to study the genetics underlying CKD and ESRD in an African population is that Africans are more genetically diverse and have the highest levels of genetic and phenotypic variation among all humans (Campbell and Tishkoff, 2008). As human populations migrated out of Africa, they carried with them part, but not all, of the ancestral genetic variation. As a result, genetic variation seen outside Africa tends to be just a subset of the genetic variants seen in Africa (Mboowa, 2014). Therefore, genetic diversity or heterogeneity is higher in Africa than the rest of the world (Gomez et al., 2014). Though Africa is more diverse genetically than the rest of the world, only few studies have been undertaken to infer genomic risk factors associated with disease in Africa (Mboowa, 2014).

Characterizing human genetic variation and examining phenotypic variation in African populations is fundamental to the identification of genes that play a key role in function and disease susceptibility (Campbell and Tishkoff, 2008). The long demographic history and variability within and between African population means that there is more genetic variation to analyse in Africans than for example European populations (Mboowa, 2014). For instance, many Europeans may share a disease associated variant irrespective of where they are from (Campbell and Tishkoff, 2008). In contrast, the frequency of variation associated with a disease in Africans may depend on the country and ethnic group of an individual.

When investigating the genetic basis of diseases, conserved variation seen in European populations mean that it is easier to identify genetic variants associated with disease risk or protection than in African populations (Gomez et al., 2014). Also, European populations are genetically very similar. To date, for other populations, it has been relatively straightforward to combine data from different studies to gain a large enough data set to perform powerful meta-analyses (Gomez et al., 2014). Thus, diversity both within and between African populations means that combining data from studies of these populations is more difficult. Therefore, the rich genomic diversity in African populations can offer new insights about disease susceptibility that could easily be overlooked using less-tailored analyses (Gomez et al., 2014).

Although Africa is critical for understanding genetic risk factors associated with diseases, it has been under-represented in human genetic studies (Gomez et al., 2014). That is why in this study I have sort to understand the genetics underlying CKD/ESRD based on an African population of mixed ancestry, targeting to harness and tap into this rich vein of genomic diversity and identify plausible genetic variation exclusively to African patients that may explain their ESRD.

1.6 DNA sequencing

Determining the DNA sequence is the most comprehensive way of attaining information about the genome of any living organism. Nucleic acid sequencing is a way to determine the exact order of the DNA bases (Lander, 2011; Venter et al., 2001). Over the past decade, the usage of nucleic acid sequencing has become accessible for researchers. Massively parallel DNA sequencing platforms have become widely available, more than halving the cost of DNA sequencing and transforming the field by putting the sequencing capacity of major genome centers into the hands of individual investigators. These new technologies are rapidly evolving, and some of their challenges include the development of robust protocols for generating sequencing libraries, building effective new approaches to data-analysis, and reconsidering and modifying experimental design. Next-generation DNA sequencing has the potential to radically accelerate biological and biomedical research, by enabling the comprehensive analysis of genomes, transcriptomes and to become inexpensive, and reducing their demanding significant production-scale efforts (Consortium and others, 2012). Also, DNA genome wide sequencing allows us to generate new hypotheses from genome wide data, whereas previously genetics studies were largely restricted to hypothesis testing.

1.6.1 First generation sequencing technology

Sanger sequencing method had become the gold standard for 30 years after its discovery in 1977. Sanger sequencing was used to obtain the first consensus sequence of the human genome in 2001 and the first individual human diploid sequence (Lander, 2011; Venter et al., 2001). This method uses DNA polymerase which makes

use of inhibitors that terminate the newly synthesized chains at specific residues. DNA to be sequenced can be prepared in two different ways, shotgun de novo sequencing or targeted re-sequencing. The output of both methods is an amplified template. Then, template denaturation, primer annealing, and primer extension are performed in cycle sequencing. With the help of fluorescently labeled ddNTPS, each round of primer extension is halted. Labeled ddNTPs in its current form are mixed with regular, non-labeled, and non-terminating nucleotides in a cycle sequencing reaction. The label on the terminating ddNTP of any fragment corresponds to the nucleotide identifying its terminal position. To separate sequences by length and to provide subsequent interrogation of the terminating base capillary electrophoresis is applied. The sequence is determined by high-resolution electrophoretic separation of the single-stranded, end-labeled extension products in a capillary based polymer gel. Laser excitation of fluorescent labels as fragments of discreet lengths exit the capillary, coupled to four-color detection of emission spectra, and provides the readout that is represented in a Sanger sequencing 'trace' (Figure 1.2). Software translates these traces into DNA sequence, while also generating error probabilities for each base-call (Metzker, 2010; Pettersson et al., 2009)

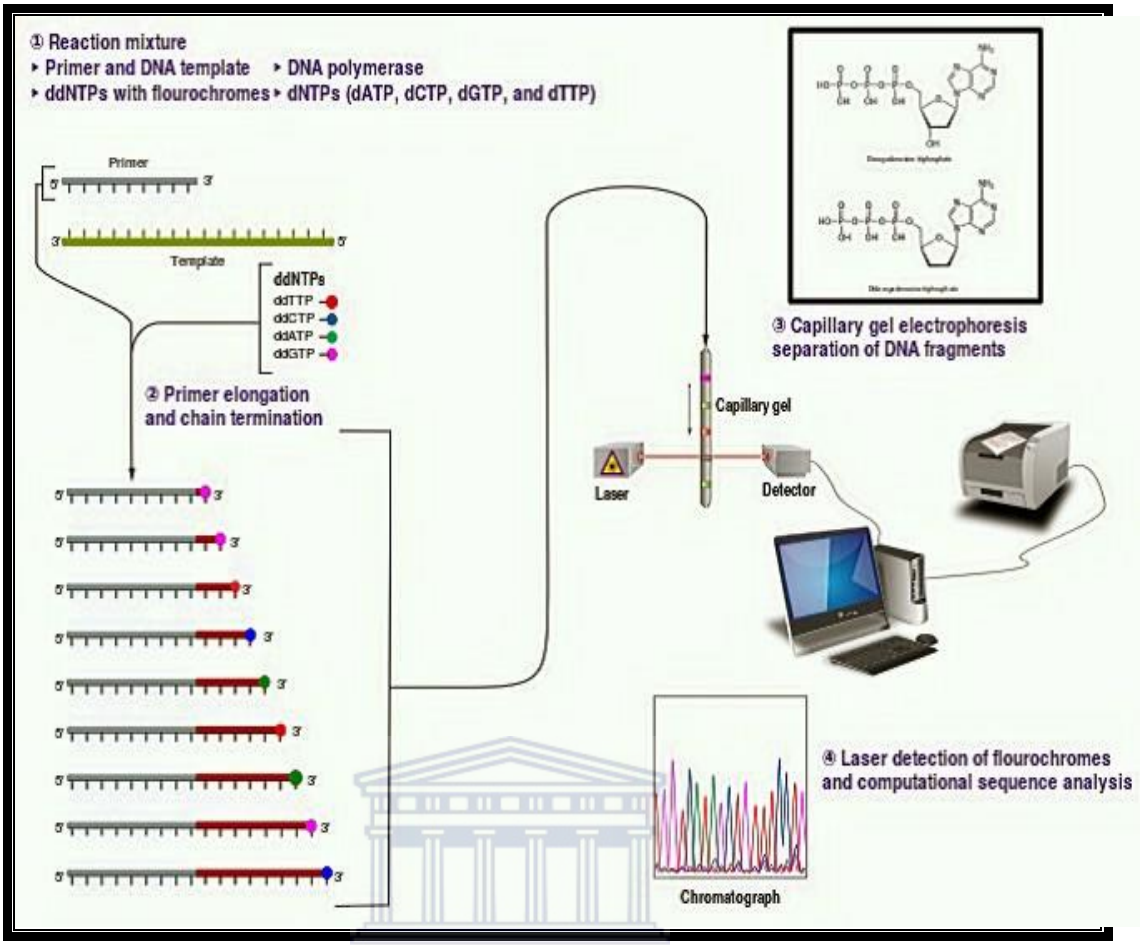


Figure 1.2: Workflow of the Sanger Sequencing method (Estevezj, 2012).

1.6.2 Next generation sequencing

After the accomplishment of the Human Genome Project, cheaper and faster sequencing methods were required in the field of biomedical research. This demand led to the development of next-generation sequencing (NGS) methods also known as massively parallel sequencing methods. NGS instruments provide higher throughput at an unprecedented speed by sequencing millions of short DNA fragments in parallel, and have become a fast, affordable approach to determine the underlying genetic causes of diseases (Metzker, 2010). Millions of fragments of DNA from a single sample can be sequenced in unison with NGS. With this technology an entire genome can be sequenced in less than ten days. In addition, the cost required for a whole human genome has decreased significantly with the use of NGS technology. It has also minimized the need for the fragment-cloning methods which are frequently used in Sanger sequencing. Currently, the three most commonly used platforms are Roche

454 (introduced in 2005), Illumina (launched in 2006) and ABI SOLiD (followed in 2008) (Grada and Weinbrecht, 2013). All three platforms sequence DNA by measuring and analyzing signals, which are emitted during the creation of the second DNA strand but differ in how the second strand is generated. In order to produce detectable signals, template DNA is fragmented into small pieces, amplified and immobilized on a glass slide before sequencing (Figure 1.2).

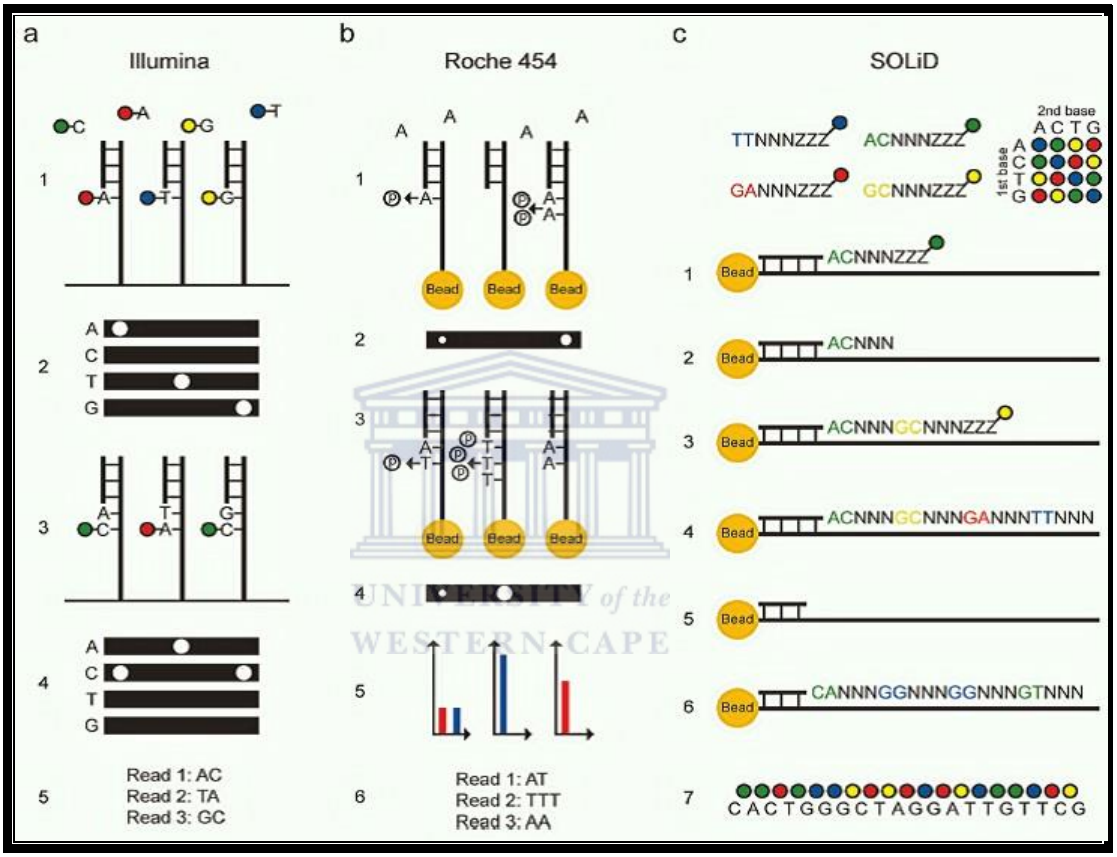


Figure 1.3 Principles of sequencing and imagin (Ansorge, 2009). Three different sequencing platforms are looked at. Illumina shows how single different bases are detected as they are added one after the other. This is done for each cluster amplified. Roche 454 shows how bases are added through sequencing by synthesis leading to the formation of the reads. SOLid shows fluorescently labelled di-base probes compete for ligation to the sequencing primer to produce the reads.

1.6.2.1 Roche 454

The first NGS technology was 454 pyrosequencing (Figure 1.4), which was also the first massively parallel sequencing technology to sequence a complete human genome, that of Dr. James D. Watson (Wheeler et al., 2008). Roche 454 implements pyrosequencing, which measures released pyrophosphates allowing the analysis of

read fragments up to a few hundred base pairs. This method is based on the "sequencing by synthesis principle" which means taking the single stranded DNA to be sequenced and sequencing its complementary strand in an enzymatic way. Using this method the activity of DNA polymerase is monitored by another enzyme, chemiluminescene. When the complementary nucleotide is bound by the single-stranded sequenced DNA, light is emitted. Sequencing is accomplished by the produced chemiluminescent signals. Since this technique infers the number of incorporated nucleotides from the signal's intensity, the system experiences problems when homopolymer stretches longer than 8 bp are sequenced (Grada and Weinbrecht, 2013). This complicates identification of small insertions and deletions in same stretches of the same DNA template.

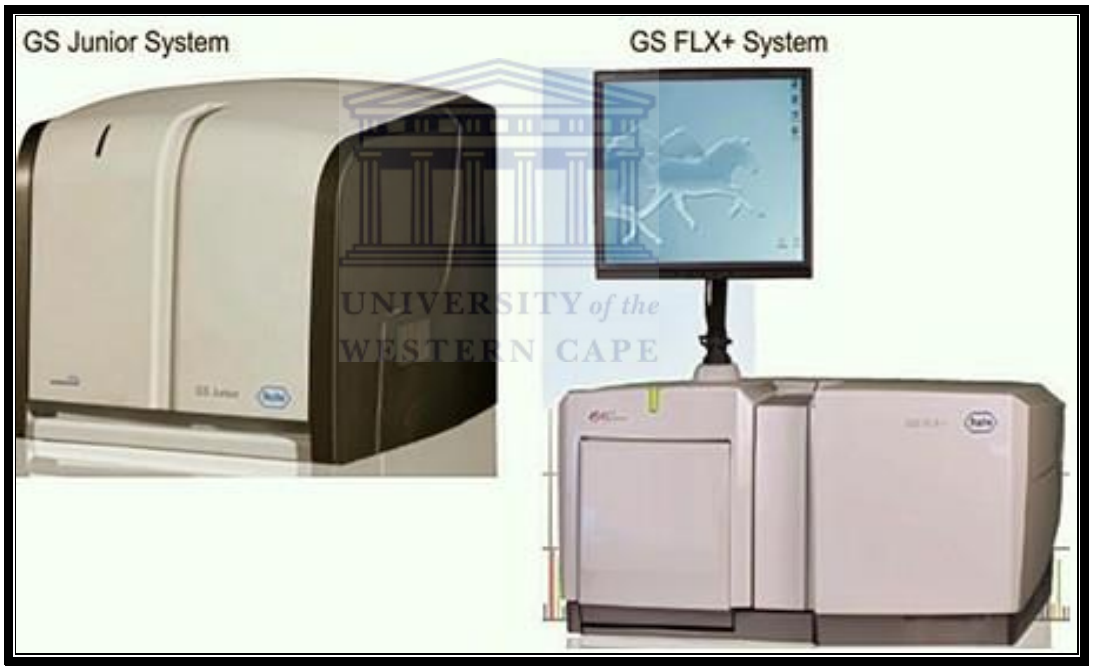


Figure 1.4 Roche 454 machine (Ansorge, 2009). The machine implements pyrosequencing, which measures released pyrophosphates allowing the analysis of read fragments up to a few hundred base pairs.

1.6.2.2 Illumina

DNA sequence data used in this project is produced using Illumina machines and protocol. Illumina's sequencing platform uses sequencing by synthesis (SBS) technology to generate exome data. The technology is able to detect single bases as

they are added to DNA strands, using a reversible terminator-based method. The fluorescent terminator is imaged as deoxyribonucleotide triphosphate (dNTP) is added, and then cleaved so that the next base can be added and imaged. Incorporation bias is minimized by competition, as all four reversible terminator-bound dNTPs are present during each sequencing cycle. SBS supports both single read and paired end libraries. The platform combines short-insert paired-end capabilities as well as long-insert paired-end reads to fully characterize the genome being sequenced. Illumina avoids homopolymer calling problems at the cost of being capable of sequencing only shorter fragments (Grada and Weinbrecht, 2013).

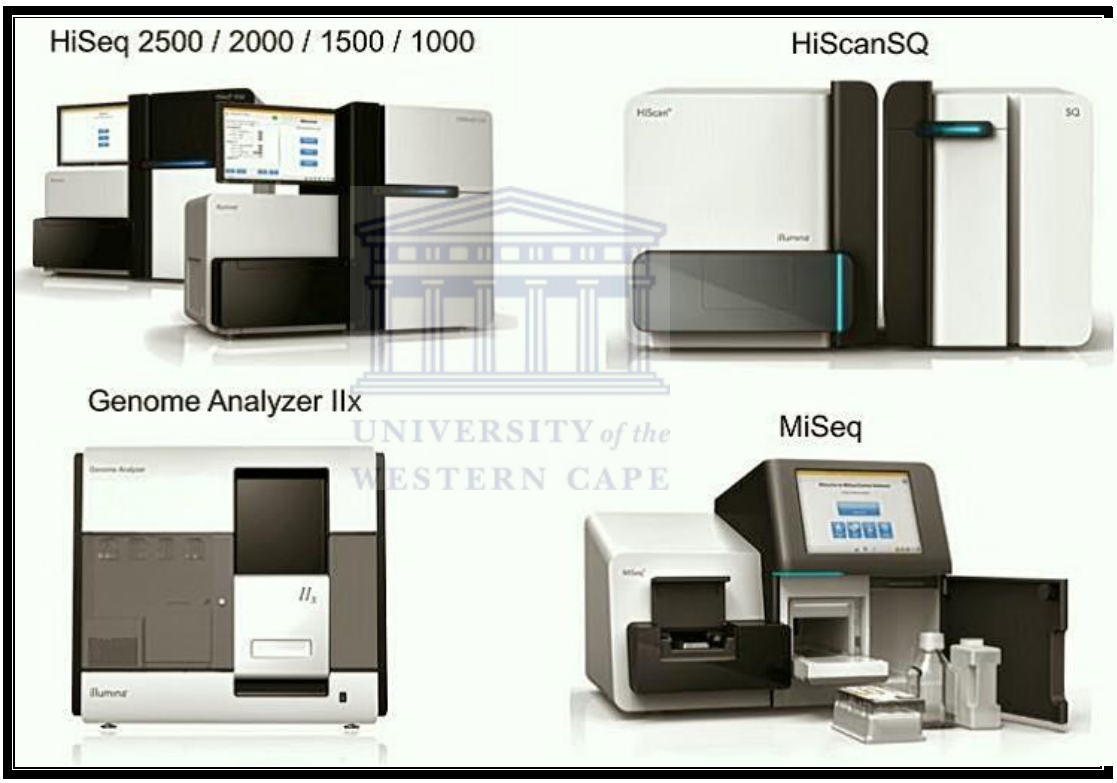


Figure 1.5: Different Illumina machines (Ansorge, 2009). Illumina sequencing machines use sequencing by synthesis (SBS) technology to generate genomic sequence data. The technology is able to detect single bases as they are added to DNA strands, using a reversible terminator-based method.

1.6.2.3 SOLiD

SOLiD sequencing platform stands for Sequencing by Oligonucleotide Ligation and Detection. Four fluorescently labelled di-base probes compete for ligation to the sequencing primer. Specificity for the di-base probe is done by interrogating every first and second base in each ligation reaction, and the eventual read length is

determined over multiple rounds of ligation, detection, and cleavage. Following a series of these ligation cycles, the extension product is removed and the template is reset with a primer complementary to the n-1 position for a second round of ligation cycles. Five rounds of primer reset are completed for each sequence tag. This allows nearly every base to be queried in two different ligation reactions by two different primers, improving the accuracy of nucleotide base calls. Variations from the reference sequence display as a fluorescent color change; sequencing errors would therefore show as one change while accurate calls would show two. Due to the nature of this approach, identified calls are not stored in nucleotide but in color space a property that needs to be considered in downstream analyses (Magi et al., 2010).

Table 1.1: A comparison of NGS sequencing technologies (Ansorge, 2009). PacBio has the capability to produce longer reads and faster but it's prone to a high error rate. Illumina paired end reads are currently the most used for exome sequencing data.

Machine	Capacity	Speed	Read Length	Cost Per Base (€)
454 Roche	35-700 Mb	10-23 hours	400-700 by	714/14285 x 10^{-8}
SOLiD	90-180 Gb	7-12 days	75 by	3/5 x 10^{-8}
Illumina	6-600 Gb	2-14 days	100-250 by	2/333 x 10^{-8}
Ion Torrent	20 Mb-1 Gb	4-5 hours	200 by	100/10000 x 10^{-8}
PacBio	1 Gb	30 minutes	3,000 by	60/80 x 10-8

1.6.3 Whole exome sequencing

One of the major endeavors of biomedical science is discovering the causal gene variants underlying human diseases. Previously, single-gene disorders were first analyzed based on Linkage analyses followed by positional cloning (Bamshad et al., 2012). Homozygosity mapping was used to ascertain loci of autosomal recessive disorders (Hamosh et al., 2005). More complex forms of single-gene disorders, such as retinitis pigmentosa and hearing loss, with different inheritance modes have been reported based on SNP arrays (Grada and Weinbrecht, 2013; Pettersson et al., 2009).

Despite this, there are still many rare genetic diseases for which causative genes remain unknown. There are also many people who may have genetic conditions but remain medical mysteries; not only that, but there are also patients receiving suboptimal treatment, or with incorrect diagnoses, yet traditional methods have been

unable to confirm the genetic cause of their disease. Drawbacks and limitations of these approaches that have hindered disease gene discovery need to be highlighted; there are families with small number of affected individuals, which do not meet the criteria required for classical gene-discovery methods. In addition, finding the causal genes in families fitting the criteria is very difficult in case of expression variability, locus and phenotypic heterogeneity, reduced penetrance or reduced fitness, because in these conditions, the causal effect could hardly be co-segregated with affected status within the family. With the advent of next-generation sequencing (NGS) technology, identification of genetic variations that underpin disease causality may be possible despite these obstacles.

Indeed, WES using NGS technologies convey novel insights into unraveling the genetic basis of diseases. The exome is the protein-coding region constituting approximately 1% of the human genome, or 30 megabases (Mb), fragmented across approximately 180,000 exons (Ng et al., 2009b). WES involves selective capture and sequencing of these protein coding regions of the genome. WES using next-generation DNA sequencing platforms has become widely available, reducing the cost of DNA sequencing by four orders of magnitude relative to Sanger sequencing (Grada and Weinbrecht, 2013; Pettersson et al., 2009). It is estimated that approximately 85% of known variation or mutations underlying disease traits occur in the exons (Choi et al., 2009b). Most of these functional variants include nonsense/missense variants, small insertion/deletions, splicing and regulatory mutations. As such, the exome represents a highly enriched subset of the genome in which to search for variants with large effect sizes which may be fundamental to unlocking the genetic basis underlying several human traits.

Since the publication of the first WES proof-of-concept report in 2009, the discovery of disease-causing genes using WES has increased rapidly, with a marked jump from 2011 to 2012 (Figure 1.6) (Ng et al., 2009b). Identification of disease causing genes may accelerate drug development by discovering disease pathways to target for new and effective treatments. In addition to gene identification, exome sequencing has also been used to correctly diagnose individuals who previously had no diagnosis or had a misdiagnosis. Also, targeted sequencing of the exome has the capacity to uncover causative genes in common disorders with complex genetic factors like cancer,

diabetes or Alzheimers disease (Chahrour et al., 2012). Sequencing of the exome also increases our current understanding of known medical disorders, like discovering relatedness of symptoms that were never before thought to be related, or discovering new information about metabolic pathways related to disease (Gibson et al., 2013). In particular, WES has been applied in characterisation of mutations in inherited disorders and rare syndromes (Gibson et al., 2013), understanding complex genetic disorders and disease risk, for example in autism, epilepsy and HTN (Chahrour et al., 2012). WES has also been successfully applied in identifying cancer driver mutations, its diagnosis as well as treatment.

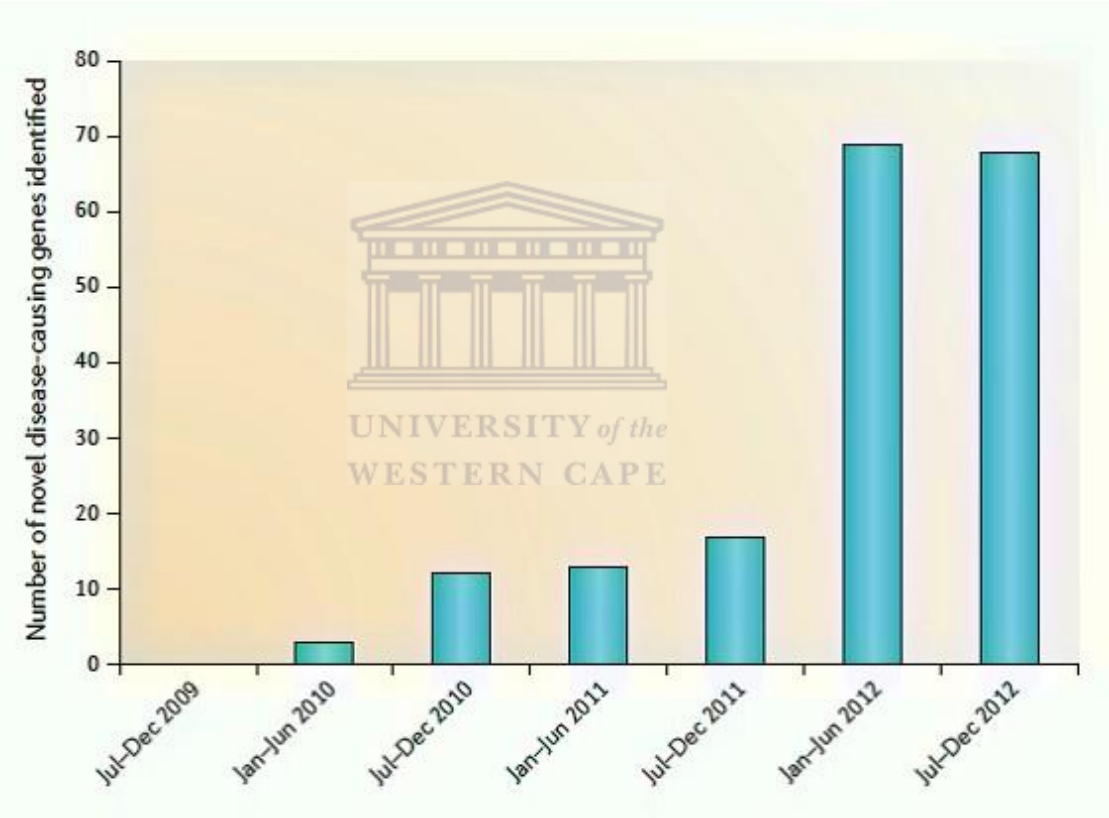


Figure 1.6: Pace of discovery of novel rare-disease-causing genes using whole-exome sequencing. Since the first WES proof-of-concept experiment in 2009, the discovery of disease-causing genes using WES has increased rapidly, with a marked jump from 2011 to 2012 (Boycott et al., 2013).

1.6.3.1 Challenges and limitations of exome sequencing

One major limitation of exome sequencing is that it only analyzes approximately 1% of the entire genome (Bamshad et al., 2012). In addition, applying exome sequencing has been hampered by how best to define the set of targets that constitute the exome.

Considerable uncertainty remains regarding which sequences of the human genome are truly protein coding, as our knowledge of all truly protein-coding exons in the genome is still incomplete. As a result current capture probes can only target exons that have been identified thus far and exclude exons not yet identified. Also, WES misses intronic sequences, structural and DNA sequence variants that may have vital regulatory functions. In other words, these are areas which could affect gene function, and ultimately disease symptoms even though they may not code for proteins. Exome sequencing may also not be ideal for understanding structural variation in genomes despite numerous algorithms which apply read depth or read pair approach attempting to resolve this problem (Krumm et al., 2012).

Other limitations and challenges with exome sequencing include genetic heterogeneity where several genes are associated with the same disorder, duplicated sequences throughout the genome and possible inadequate coverage of the gene sequence. When using exome sequencing in clinical genetics and medicine, limitation of the approach is evident and experimental design is needed to circumvent the problem. Genetic and phenotypic heterogeneity in different affected individuals can make exome sequencing data difficult to interpret. Patients with the same phenotype may not share the same causal variant; actually they may have distinct variants in a gene, in what is referred to as allelic heterogeneity. Intensive analyses of variant calls are important in exome sequencing. False-positive errors appear as sequencing errors related to mechanical and analytical errors. Also short reads generated by NGS would not align perfectly to the appropriate position as a result of paralogous and low copy repeats that may cause errors during calling (Liu and Leal, 2010). Some deleterious variations may be located in the non-coding regions, such as intronic or regulatory regions of the genome, which cannot be called by exome sequencing, or may be located in the region of the genome that may not be adequately covered.

1.6.4 Whole exome sequencing analysis workflow

Comprehensive NGS data analysis process is multifaceted, it comprises manifold analysis steps, and the process is reliant on a multitude of programs, databases and involves handling enormous amounts of heterogeneous data (Figure 1.7). It is

unsurprising that due to the enormous success of NGS projects, a deluge of tools have been created to support specific parts of the analysis workflow. It is apparent that the appropriate choice of tools is a not a trivial task and is dependent on several factors.

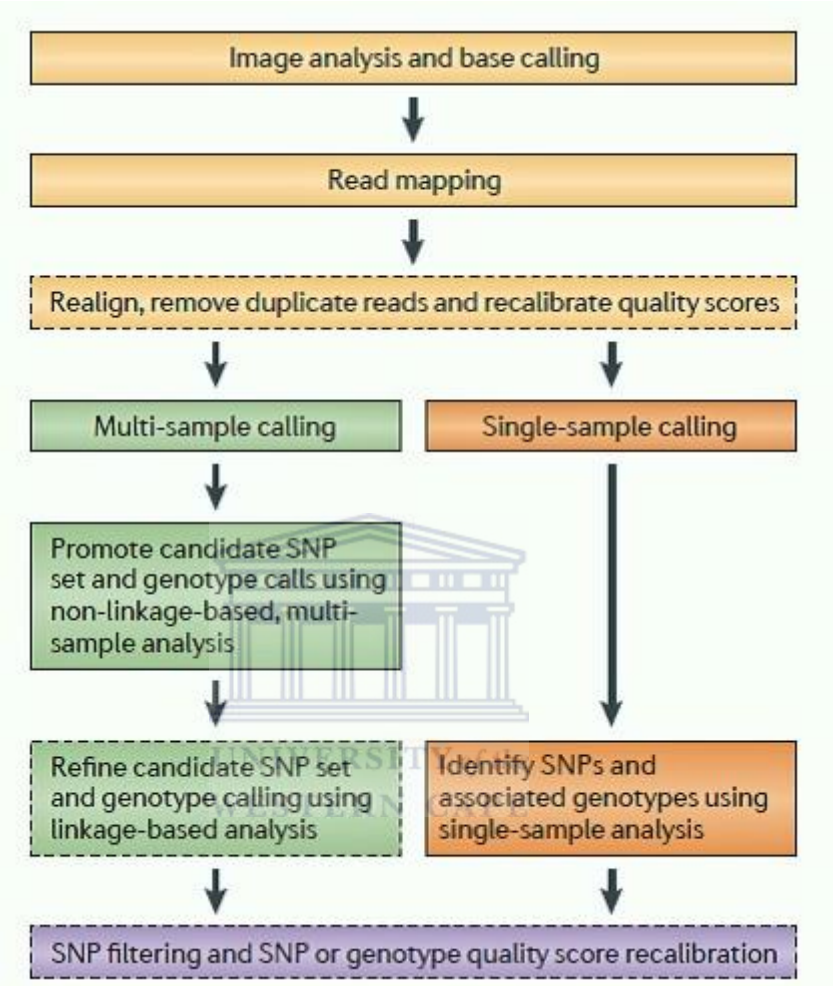


Figure 1.7: Basic protocol for whole-exome sequencing data analysis. Image analysis and base calling is usually done at the core sequencing facilities. The reads produced are mapped to the human reference genome using various alignment softwares. The alignments are refined further by performing local realignment and recalibration of quality scores. Variant calling is undertaken on the aligned reads and variants are produced. The variants are also recalibrated to reduce errors

1.6.4.1 Library preparation

Appropriate material for sequence analysis is indispensable. This requires correct identification of the proband, making appropriate clinical diagnosis and the sample must be collected, identified, recorded and stored under quality controlled conditions suitable for diagnostic testing. For example, if a case has been identified as part of a research project it may be necessary to collect an additional sample. Genomic DNA

from peripheral white blood cells is the typical starting material.

The major NGS platforms are Illumina, Roche, Solid and Agilent. Each of these platforms is compatible with the main commercial options for the first step of WES, which is enriching the exonic sequences. The sequencing platform kits tend to contain exons from the consensus coding sequence project, which currently comprises approximately 180,000 exons from roughly 18,409 genes, as well as additional sequences (Robinson et al., 2011). Each company also has developed its own exome enrichment platform (Agilent's SureSelect Human All Exon 50Mb, Roche/Nimblegen's SeqCap EZ Exome Library v.2.0, and Illumina's TruSeq Exome Enrichment), which differ in design and experimental parameters that can affect variant discovery (Table 1.2). Clark et al. 2011 performed a systematic analysis of their differences (Clark et al., 2011).

Nimblegen uses DNA for capturing targeted genomic sequences. The platform contains overlapping DNA baits that cover target bases multiple times, resulting in the highest density coverage of the three platforms. It covers a greater portion of miRNAs compared to other enrichment platforms. Agilent uses RNA for capture of targeted genomic sequences, where RNA baits reside immediately adjacent to each other across target exon intervals. It provides better coverage of genes in the Ensembl database; Whilst Illumina uses DNA for capture of targeted genomic sequences and relies on paired-end reads to extend outside bait sequences and fill gaps. The majority of targets unique to this platform cover untranslated regions (UTRs).

Table 1.2 Details of human exome capture techniques. The average targeted region for the exome is approximately 50Mb (50 million base pairs) covering an average of 180000 exons. This targeted region enables one to interrogate more than 20000 genes simultaneously

	Illumina	Agilent sureSelect			Roche Nimblegen	
	Exome	Human all Exon 50Mb	Human all Exon V4	Human all Exon V4 +UTRs	Version 2.0	Version 3.0
Targeted region size	62 Mb	50 Mb	51 Mb	71 Mb	36.5Mb	64Mb
Number of target genes	20,794	20,718	20,956	20,965	30,000	24,685
Number of target exons	201,121	331,518	334,378	335,765	300,000	220,000

1.6.4.2 Base calling and quality control

In addition to sequence data, base calling produces quality scores for each base, which are estimates of the probability of the call being erroneous. After base calling, reads with indications of varied signals or other errors are filtered out. In addition, reads that do not start with a specific key sequence, which is part of the adapter, and reads which have a high number of off-peak signal intensities (indicative of homopolymer errors) are filtered out. This is a vital step that enables the discovery of anomalies that may have originated from the sequencer or library material used during the sequencing process, thereby averting error from proliferating and generating false variant calls in subsequent analysis. It also affords an opportunity to recover valuable data by trimming off the poor quality segments that generally occur at the 3'end of reads. Sequence reads are trimmed from the 3' end primarily to remove adapter sequence and bases of low quality, which may have ascended from phasing/prephasing issues and loss of signal intensity. A common, broadly used tool for assessing the quality of sequence data is the FastQC software which reports distributions of base qualities, GC content, redundancy and over-representation of adapter or primer sequence (Bioinformatics, 2011).

1.6.4.3 Read mapping and Alignment to reference genome

Alignment is the process of mapping short nucleotide reads to a reference genome (Figure 1.8). Each of the millions of short reads must be matched to the 3 billion possible positions within the human genome. This is a critical computational step for downstream analysis, especially variant calling. Different mapping tools assess the likely starting point of each read within the reference genome, a process which is complicated by the volume of short reads, unique versus non-unique mapping, and variation in base quality. This is a computationally challenging and time consuming undertaking (Day-Williams and Zeggini, 2011). It is also a vital step, as any errors in alignment to the reference genome will be carried through the rest of the analysis. The Sequence Alignment/Map (SAM) and Binary Alignment/ Map (BAM) formats are the standard file formats for storing NGS read alignments (R. Li et al., 2009).

Short reads generated from NGS may both be single end reads or paired-end reads and vary from dozens to hundreds of base pairs (Ruffalo et al., 2011). These reads need to be aligned correctly to their appropriate location within the reference genome. The task is complicated by many factors, which include genetic variation in the population, sequencing error, short read length and the huge volume of short reads to be mapped. To date, numerous algorithms have been developed to overcome these challenges and have been made available to the scientific community as software packages.

Alignment programs have different properties in terms of their ability to perform gapped alignment, how base qualities are used during alignment and how reads aligning to repeated regions are treated. Some aligners can handle data from any sequencing platform, whereas others are specific to one platform. To save computational time, the alignment is typically executed in two steps: a limited number of candidate positions are identified by fast heuristic approaches, and candidate positions are evaluated by more accurate methods, such as the Smith-Waterman algorithm (Li and Homer, 2010). Currently some of the available software packages for short read alignment include Bowtie (Langmead and Salzberg, 2012), SOAP (R. Li et al., 2009), BWA (Li and Durbin, 2009), and Novoalign (Novocraft (2010), http://www.novocraft.com/). Bowtie, BWA and SOAP align quickly but require

significant amounts of time to build an index of a genome. Novoalign, conversely, requires little indexing time and it has become quite popular in recent studies due to its accuracy and is a preferred aligner used in this project (Ruffalo et al., 2011).

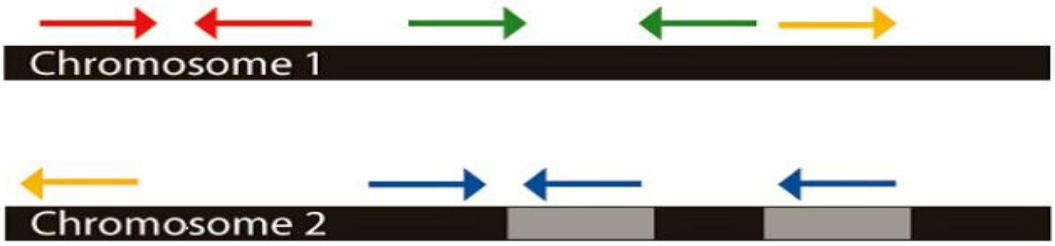


Figure 1.8 Principles of reference alignment of paired-end reads to a reference genome. Arrows with the same color indicate reads that belong to the same pair. Red arrows illustrate a normal pair, aligning with the expected orientation and distance. Green arrows illustrate a pair that aligns at a larger distance than expected due to a potential deletion in the sequenced genome. Orange arrows illustrate a pair that aligns to different chromosomes indicating a potential rearrangement in the sequenced genome. Blue arrows illustrate how paired-end reads can guide alignment if one of the reads aligns in a repeated (grey) region. The correctly aligned read will provide the correct alignment position for the misaligned read pair.

1.6.4.4 Variant calling and genotyping

Variant calling is the subsequent step taken after alignment of reads to the reference sequence. Since the reads are already aligned, the sample genome is compared to the reference genome so that variants can be identified. These variants may be responsible for disease, or they may simply be genomic noise without any functional effect or affecting non etiological characteristics. Variant call format (VCF) is the standardized generic format for storing sequence variation including SNPs, indels, larger structural variants and annotations (Consortium and others, 2010; Danecek et al., 2011; Via García et al., 2012). The major computational challenges in variant calling are due to the issues in identifying “true” variants as opposed to alignment and/or sequencing errors. Yet, the ability to detect SNPs with both high sensitivity and specificity is a vital step in identifying sequence variants associated with disease, detection of rare variants, and assessment of allele frequencies in populations.

Variant calling is complicated by three factors: (1) the presence of indels, which represent a major source of false positive SNV identifications, especially if alignment algorithms do not perform gapped alignments, (2) errors from library preparation due

to PCR artifacts and (3) variable GC content in the short reads (Consortium and others, 2010; Danecek et al., 2011; Via García et al., 2012). Thus, the rate of false positive and false negative calls of SNVs and indels is a concern. It is therefore recommended that before the variant calling process is carried out additional quality control steps should be done. These include alignment refinement to improve the accuracy of the data by local realignment, removing PCR duplicates, and recalibrating variant quality scores (Consortium and others, 2010; Danecek et al., 2011; Via García et al., 2012). Even though these steps go a long way towards improving accuracy of variant calling other additional steps such as variant recalibration are also recommended (Consortium and others, 2010; Danecek et al., 2011; Via García et al., 2012).

One extensively used tool which utilizes most of these quality control steps is the Genome Analysis Toolkit (GATK) (McKenna et al., 2010). Developed by the Broad Institute, the Genome Analysis Toolkit (GATK) is one of the most popular methods for variant calling using aligned reads. It is designed in a modular way and is based on the MapReduce functional programming approach (McKenna et al., 2010). GATK may be used for single or multi sample variant calling for exome sequencing data. The package has been used for projects such as The Cancer Genome Atlas and the 1000 Genomes Project (Consortium and others, 2010; Network and others, 2012). Other tools that are also useful for variant calling are CRISP (Bansal, 2010), SAMtools (R. Li et al., 2009), SNVer (Wei et al., 2011) and VarScan (Koboldt et al., 2012).

1.6.4.5 Variant annotation

After alignment and variant calling, a list of thousands of potential differences between the genome under study and the reference genome is generated. The next step is to determine which of these variants are likely to contribute to the pathological process under study. Many tools exist to examine relevant variants by referencing previously known information about their biological functions and inferring potential effects based on their genomic context. There are several tools available for use in this step and the ones that have gained wider use are:

ANNOVAR is a tool used to perform up to date functional annotation of various genomes, supporting SNPs, INDELs, block substitutions as well as copy number

variants (CNVs). The tool offers a wide selection of diverse annotation techniques, structured in disparate categories such as gene-based, region-based and filter-based annotation. The tool depends on several databases, which need to be downloaded individually and updated constantly. The SeattleSeq Annotation server (http://snp.gs.washington.edu/SeattleSeqAnnotation) offers a web application for annotating human SNPs and INDELs. In contrast to most other web-based annotation services, SeattleSeq affords the opportunity to directly upload input files in various formats for batch analysis of multiple variants. As variant annotation is performed on a remote server, the tool might be helpful for research groups without committed hardware for data analysis.

snpEff (Ge et al., 2011) is a popular variant annotation tool, which has also been integrated within Galaxy and GATK. In addition to SNPs, the package also supports analysis of INDELs and multiple-nucleotide polymorphisms. snpEff identifies numerous diverse effects, which are categorized into four classes (high, moderate, low and modifier) by their putative functional impact.
Variant effect predictor (VEP) (Medina et al., 2012) is Ensembl's own functional annotation tool, previously known as SNP effect predictor. The tool can be used either by a web interface, as command-line tool or via a Perl API. The web interface version is aimed at users analyzing smaller sets of variants, as it is only capable of processing approximately 750 variants per file.

All annotation tools provide a set of broad attributes for each recognized mutation. These properties can be used to evaluate the plausible impact of each mutation. All tested applications provide links to one or more public databases of known mutations. ANNOVAR uses six different scores: GERP (Davydov et al., 2010), LRT (Chun and Fay, 2009), MutationTaster (Schwarz et al., 2010), PolyPhen (Adzhubei et al., 2013), PhyloP conservation (Pollard et al., 2010) and SIFT (Kumar, 2013). SeattleSeq supplies four scores: GERP, Grantham, phastCons (Langmead and Salzberg, 2012) and PolyPhen. NGS–SNP and VEP provide three scores: Condel (Gonzalez-Angulo et al., 2010), PolyPhen and SIFT. These scores are computed based on various different approaches, such as sequence homology, evolutionary conservation, protein structure or statistical prediction based on known mutations.

1.6.4.6 Statistical prioritization and candidate gene identification

The advance of exome and genomic sequencing is yielding new information about an extensive number of human genetic variants. A number of candidates disease-associated SNVs can be identified following alignment and variant calling. Unlike nonsense and frameshift mutations, which often result in a loss of protein function, pinpointing disease-causal variants among numerous SNVs has become one of the major challenges. For instance, approximately 1,300 loci are shown to be associated with roughly 200 diseases by GWAS but only a few of these loci have been identified as disease-causing (Lander, 2011). Exome sequencing enables the identification of more novel genetic variants than previously imaginable, but it still requires computational and experimental approaches to predict whether a variant is deleterious. To this end, several approaches have been developed to identify rare, non-synonymous SNPs that cause amino acid substitution (AAS) in the coding region. These include SIFT, PolyPhen, GERP and MutationTaster (Schwarz et al., 2010) amongst others.

However, a challenge that remains is one of narrowing down the list of candidate variants and interpreting retained variants within a biological context (De Baets et al., 2011). A widely used approach to substantially reduce the candidate list is to exclude known variants which are present in public SNP databases, published studies or in-house databases as it is assumed that common variants represent harmless variations (Wang et al., 2010). Various tools are used for variant prioritization and candidate gene identification such as The Variant Analysis Tool (VAT), The Variant Annotation, Analysis and Search Tool (VAAST) and Ingenuity variant analyzer (IVA). In this project VAAST and IVA were used. VAAST identifies damaged genes and deleterious variants in personal genome sequences using a probabilistic search method (Yandell et al., 2011). The tool utilizes both existing amino acid substitution and aggregative approaches to variant prioritization and combines them into a single unified likelihood framework. The method increases the accuracy with which disease causing variants are identified. VAAST scores coding and noncoding, rare and common variants simultaneously and aggregates this information to identify disease causing variants. IVA ccombines analytical tools and integrated content to help rapidly identify and prioritize variants by narrowing down to a small, targeted subset of compelling

variants based both upon published biological evidence and knowledge of disease biology. With IVA, variants can be interrogated from multiple biological perspectives; the program will explore different biological hypotheses and identify the most promising variants for follow-up.

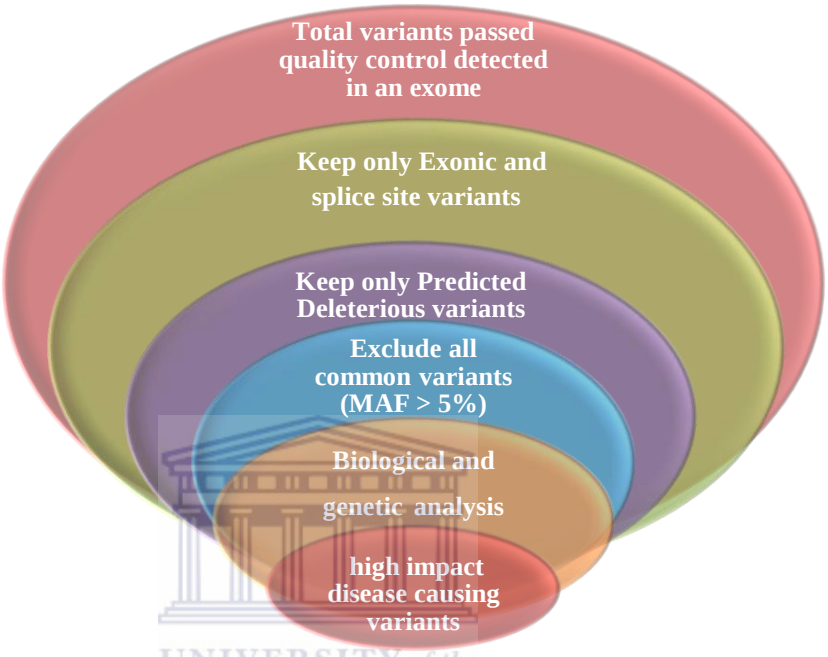


Figure 1.9 Filtering steps followed in variant prioritization of exome sequencing data. From the total variants identified in each individual exome, only splice site and exonic variants are kept, discarding intron and other variants which may be identified. Only rare variants that are expected to have an effect on the protein coding are kept. Filtering further will retain variants predicted to be involved in the disease of interest as well as considering the inheritance pattern (keep only heterozygous variants for a suspected autosomal dominant inherited disease. The remaining list will contain potentially causative variants.

1.6.4.7 Data visualization

Genome browsers (GB) can be classified into two main categories, namely web-based applications and stand-alone tools. GBs contain information about the reference genome, the transcriptome, aligned reads, found mutations, annotations collected from public data sources or other data types important for the correct interpretation of results (Loraine and Helt, 2002).

1.7 Application of whole exome sequencing in the study of disease genetics

The development of massively parallel sequencing technology has led to a dramatic acceleration in the pace of genetic discovery. NGS has improved our understanding of the genetic pathology of many diseases. Understanding the pathogenic mechanism of a disease mostly depends on finding the causative gene and variants associated with the phenotype, and analyzing the functional effect of the pathogenic variant. NGS technologies have enabled two major advances of relevance to the discovery of disease associated genes. The first is the ability to readily sequence the genome of a single person, thus allowing the identification of mutations that are specific to that individual (Figure 1.10). The second is the application of NGS to sequence the entire exome (WES), enabling the identification of mutations that result in changes to amino acid sequence of encoded proteins while substantially reducing both the cost and the computational requirements associated with analysing the resulting data. WES is also used in identifying rare, novel, as well as common genetic variants in coding regions associated within complex and common traits. Genomic regions identified by GWAS are then deep sequenced to identify causative variants.

Sequencing of human exomes was first reported by Ng et al 2009 (Ng et al., 2009b). The author reported the targeted capture and massively parallel sequencing of 12 human exomes including eight individuals previously characterized by HapMap and Human Genome Structural Variation project and four unrelated individuals affected by an inherited disorder called Freeman-Sheldon (FSS). FSS is a rare autosomal dominant disorder for which the associated gene, MYH3, was earlier identified using other methods. This study established that exome sequencing could be used to identify causes of rare disorders using few affected individuals. It also demonstrated that exome sequencing was cost-effective, reproducible and a robust approach for the identification of medically significant genetic changes. This study was effective in ascertaining causative mutation in MYH3 gene and also confirmed the utility of exome sequencing in uncovering the genetic basis of genetic disorders.

Soon after, WES was performed on four patients with Miller Syndrome. Two previously unknown variants in each of the four individuals identified a single candidate gene, DHODH. Sanger sequencing confirmed the presence of DHODH mutations in three additional families with Miller Syndrome (MS) but the mutation was absent in the matched control samples (Ng et al., 2009b). Up until that point, the gene that caused MS was completely unknown. This research emphasized the efficiency and value of exome sequencing in identifying causative genes in rare genetic disorders with clear definitive phenotypes, which typically affect only a small number of people worldwide. Similarly, using targeted exome sequencing, mutations identified implicated MLL2 as a causative gene for Kabuki syndrome (Ng et al., 2010). Sequencing of the exome was undertaken in a small family with consanguinity brain malformations of cortical development (MCD), a condition with wide spectra of symptoms. Researchers discovered two copies of the same mutation in the WDR62 gene in two affected family members. These findings strongly implicated WDR62 in the cause of cortical abnormalities (Bilgüvar et al., 2010). This research brought to the fore exome sequencing's capability to detect a single gene associated with a medical condition that has several symptoms and other genetic factors.

Exome sequencing using NGS technology was performed on two siblings who had diamond blackfan anemia (DBA). A mutation in the gene encoding the hematopoietic transcription factor GATA1 was identified. This mutation was replicated in an additional patient carrying a distinct mutation at the same splice site of the GATA1 gene. These findings provided insight into the pathogenesis of DBA (Sankaran et al., 2012).

In 2013, targeted capture and WES using NGS technology was used to resolve apparent incidental findings and revealed further complexity of SH3TC2 variant alleles causing Charcot-Marie-Tooth neuropathy. An additional SH3TC2 variant that plausibly contributes to the phenotype was identified (Lupski et al., 2013). WES identified mutations in the MYO15A gene as a plausible cause of autosomal recessive non syndromic hearing (ARNH). This was an interesting study as ARNH is a genetic heterogeneous disorder for which it is difficult to ascertain a genetic diagnosis using other methods (Woo et al., 2013). In a large pedigree with Familial Dilated Cardiomyopathy (FDC), a complex disease, application of WES identified causative

mutations in the RBM20 gene (Wells et al., 2013), illustrating further the utility of exome sequencing in the study of genetics underlying complex diseases. Targeted exome sequencing can also be used in combination with other traditional candidate gene identification methods. Thus, by using exome sequencing and linkage analysis in a five-generation Chinese family with non-syndromic Hearing Loss, novel Tenascin-C (TNC) was implicated as a causative gene (Zhao et al., 2013). Overall, understanding the genetic basis of rare and extreme phenotypes leads to a better understanding of the disease mechanism and physiology, which obviously helps families, patients and health care providers in managing and treating the disease.

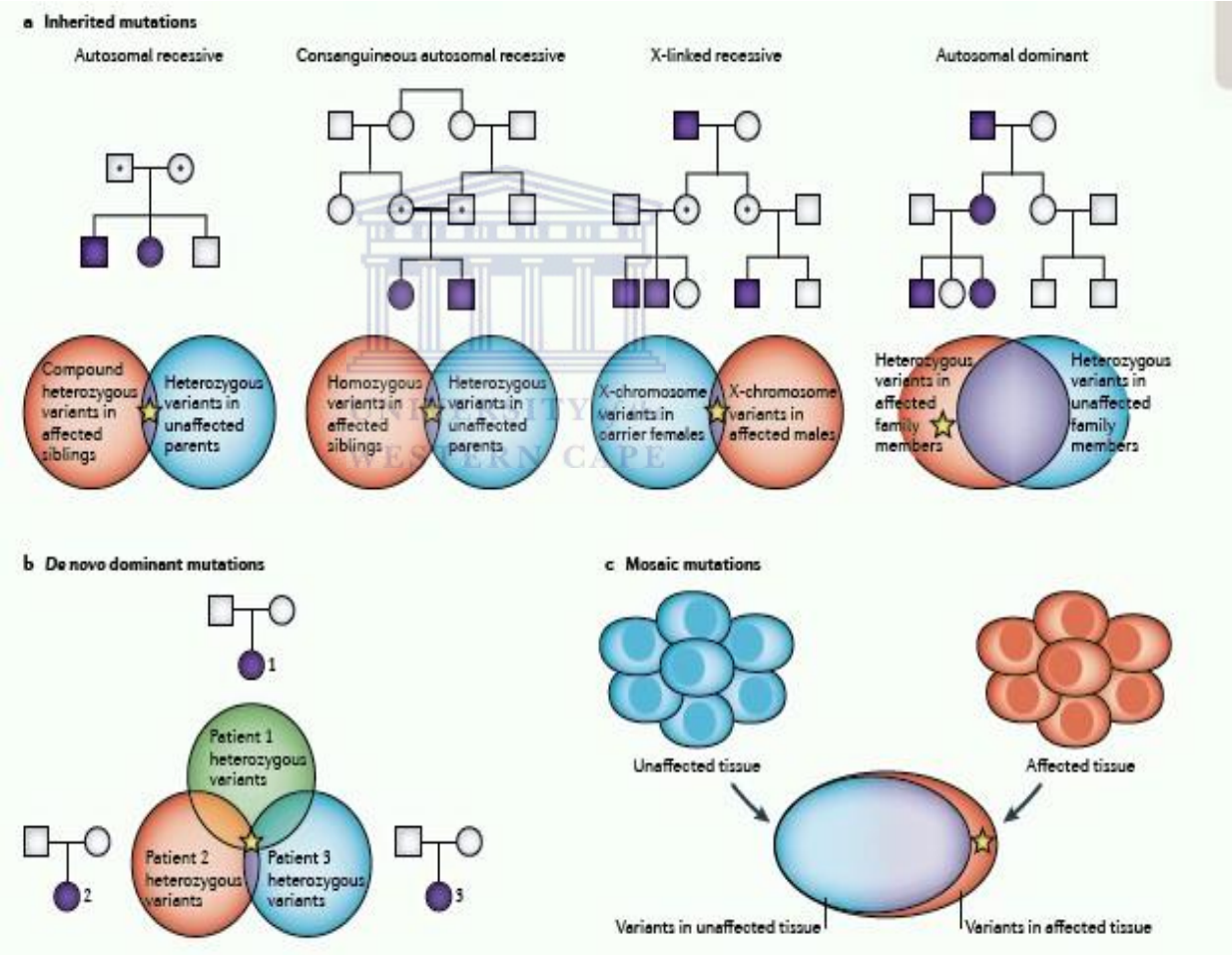


Figure 1.10 Gene identification approaches for different categories of rare diseases. Representative family structures are indicated by the pedigrees for each type of mutation (Boycott et al., 2013). The far right corner shows an autosomal dominant inherited disease and illustrates that for such a disease heterozygous variants in the affected family members that are absent in in the unaffected family members are prioritised. For an autosomal recessive inherited disease the variants sought are those that are shared by both affected and unaffected and this is usually a smaller number than those for an autosomal dominant disease

1.7.1 Whole exome sequencing as a diagnostic tool in clinical settings

The aim of WES in modern medicine is to provide an efficient and effective genetic method that can be used to implement best treatment, particularly accurate, fast and cost-effective diagnosis of the patients. The first report that pinpointed an exact diagnosis ascertained by WES was published by Choi et al. (Choi et al., 2009b). WES performed on a patient referred with Bartter syndrome, a rare inherited disorder characterized by hypokalemia. Additional analysis showed that the patient had a novel homozygous mutation in SLC26A3 gene. Mutations in this gene have been known to cause congenital chloride-losing diarrhea (CLD). Clinical re-evaluation of the patients who were misdiagnosed as Bartter syndrome determined the correct diagnosis as CLD. This study demonstrated that exome sequencing could be used to provide a correct diagnosis, alter medical management and augment current understanding of a medical disorder (Choi et al., 2009b). In another study WES helped in the diagnosis of a patient with Leber congenital amaurosis that had mutation in PEX1 gene associated with peroxisome biogenesis disorders (Majewski et al., 2011).

A 15-month-old boy with an immune deficiency disorder was exome sequenced. Analysis of the data resulted in the diagnosis of Crohn disease being made. This was owing to the identification of a mutation in the X-linked inhibitor of apoptosis gene (Worthey et al., 2010). After being properly diagnosed, he was able to receive targeted medical treatment to prevent the development of life-threatening illness, demonstrating how exome sequencing has the ability to find a correct diagnosis in an individual who has challenging to clinically disease presentation, thus making effective treatment plausible.

Again, WES was applied for neonatal screening of diabetes mellitus (NDM). It was established that patients carrying a mutation in KCNJ11 or ABCC8 genes should be given oral treatment with sulfonylurea drugs instead of insulin. In cancer studies WES identified a spectrum of mutation frequencies in advanced and lethal prostate cancers (Kumar et al., 2011).

The clinical utility of WES was also demonstrated in a study carried out by Stitziel et al. where sequencing of the exome in combination with directed clinical phenotyping was used to make a primary definitive diagnose of cholesterol ester storage disease which initially presented as Autosomal Recessive Hypercholesterolemia (Stitziel et al., 2013). Thus, making effective corrective diagnosis where primary first line clinical diagnosis would have erred. In another study, a novel frameshift homozygous variant was identified in the SACS gene, which resulted in a successful diagnosis of autosomal recessive spastic ataxia of Charlevoix-Saguenay in a single pediatric case (Liew et al., 2013). (Pangrazio et al., 2014) identified a homozygous variant in the CTSK gene, implicating it as a possible causative gene in two siblings with Autosomal Recessive Osteoporosis, initially thought to be affected by intermediate osteoporosis (Pangrazio et al., 2014). In a trio based study, WES was successfully applied to accurately diagnose 73% of children with Neurodevelopmental disorders. In the same study it was established that had WES been performed at symptom onset, genomic diagnoses may have been made 77 months earlier than the time it occurred (Soden et al., 2014).

1.7.2 Potential of exome sequencing in kidney disease genetics

Exome sequencing has been applied rapidly for variant discovery in research settings. The recent improvement in its accuracy has enabled development of clinical exome sequencing for mutation identification in patients with suspected genetic diseases. WES is a comprehensive method that can be used in the molecular diagnosis of patients with undiagnosed disorders despite having exhausted other primary diagnostics methods (Biesecker and Green, 2014; Dixon-Salazar et al., 2012; Green et al., 2013; Need et al., 2012)

In a study by Gibson et al. exome sequencing was successfully applied to resolve differential diagnosis of familial kidney disease. Initially, pathological examination was carried out on patients who presented with hematuria and proteinuria which eventually led to ESRD. A diagnosis of focal segmental glomerulosclerosis (FSGS) was made. However, when WES was applied, a mutation in COL4A5 a gene known to cause Alport syndrome was discovered. The mutation in the gene segregated with

the disease in this family. As a result, a new diagnosis of Alport syndrome was confirmed (Gibson et al., 2013).

In another study, WES was used to diagnose and distinguish cystic kidney diseases from phenocopies of renal ciliopathies in rare, genetically heterogeneous cases of CKD in children (Gee et al., 2013). Mutations in the known CKD causing genes SLC4A1 and AGXT were identified. The study established that for individuals with early onset renal failure, histological examination may represent a relatively blunt diagnostic tool, which can be incapable of establishing the correct diagnosis. Targeted exome capture of COL4A3/COL4A4/COL4A5 followed by NGS has also been used clinically to screen suspected AS patients (Fallerini et al., 2014; Morinière et al., 2014).

WES was applied in three families with unexplained inherited kidney disease. Novel COL4A3 and COL4A4 mutations were identified. The results resolved diagnostic confusion arising from atypical or incomplete clinical and histological findings which resulted in appropriate genetic counselling and treatment advice being given accordingly (Lin et al., 2014). Again, using exome sequencing rare hereditary variants in COL4A3 and COL4A4 genes were identified in patients with FSGC (Malone et al., 2014). One of the clinical significance of this discovery is that there is an overlap between phenotypes induced by COL4A3 and COL4A4 variants and familial FSGS genes. Therefore, screening for rare variants/mutations in these genes in families referred with a diagnosis of familial FSGS is warranted for better disease definition and treatment to be attained. Furthermore, exclusion of variants in these genes should be considered as part of a filtering process in the analysis of WES data in familial FSGS.

Using exome sequencing a novel COL4A5 Mutation was identified in family members who were diagnosed with Alport syndrome (AS). This discovery broadened the mutation spectrum in the COL4A5 gene associated with AS, which may also shed new light on genetic counseling for AS patients (Xiu et al., 2014). In another study, a novel mutation in the LMX1B gene was discovered in patients diagnosed with end-stage renal disease (ESRD) of unknown cause presenting in a familial autosomal dominant pattern. Such a molecular genetic diagnosis of LMX1B nephropathy may

provide a definitive diagnosis preventing the need to undertake risky procedures such as renal biopsies, and also allows family members at risk to be screened (Edwards et al., 2014). Based on the studies reviewed, the utility of exome sequencing using NGS techniques to study the genetics underlying rare and complex kidney phenotypes cannot be over emphasised. As such this approach is also adopted in this thesis.

1.8 Thesis rationale and objectives

Sub Saharan Africa (SSA) bears approximately 11% of the world population and 24% of the global disease burden but has only approximately 3% of the health work force and infrastructure (Mensah and Mayosi, 2013; Pugsley et al., 2009; Schieppati and Remuzzi, 2005; Stanifer et al., 2014). Particularly, in SSA the infrastructure for human genetics research is beyond scarce. Although lack of infrastructure that includes bio-specimen repositories is a major concern, the scant representation of SSA in human genomics research is also hugely attributed to lack of well-defined and phenotypically well-characterised disease cohorts. Such essential information may be collected and kept in well-structured and organised clinical databases. Although there are several hospital and community based epidemiological studies of kidney diseases and its risk factors (such as Lupus) from African countries, most of the studies are generally retrospective with insufficient information on risk factors, treatment parameters, important clinical outcomes and mortality. One of the major difficulties recognised for this is lack of/non-existence of clinical databases which may contain such worthwhile data. Yet, the potential of genomics research to yield valuable insights into the biology of diseases is dependent upon accurate definition of disease phenotype and a comprehensive understanding of environmental factors. In light of this, the primary aims of this bioinformatics thesis are:

(1) Develop an exome sequencing analysis pipeline using next-generation sequencing technology and apply it on a rare and difficult familial autosomal dominant form of kidney disease. The patients were identified at Groote Schuur hospital, Western Cape, South Africa. Though this is a single familial study, it illustrates the plethora of genetics studies that can be undertaken with a cohort of well-defined, phenotypically well characterised, collected and kept in a standardised clinical patient database as

alluded above. I intend to answer the following questions:

(a) What are of the genetic variants likely to be underlying kidney disease in this South African family?

(b) What gene regulatory pathways are involved in kidney disease in this family, in the context of ESRD and in the broader patient population?

(c) How can we use computational characterization of pathways and gene regulatory networks analysis using a range of computational tools, to identify fundamental similarities and differences between ESRD in the South African affected family, and the wider knowledgebase about genetics underlying ESRD?

(2) To create a holistic new paradigm which intends to increase the capacity of genomics research in SSA by developing and setting up a multicentre registry of Lupus patients (African Lupus Genetics Network), facilitating the examination of hypotheses concerning disease genetics, aetiology and health outcomes of patients. Lupus is one of the major risk factors of ESRD that is prevalent amongst CKD patients that are treated at Groote Schuur Hospital in the Western Cape, South Africa. The clinical database will enable researchers to establish a disease cohort with an ample sample size for genomic research to be embarked on and will provide a platform for other translational research of kidney disease related conditions. Access to comprehensive, standardised and precise phenotypic data from well characterised research participants is an essential complimentary tool to genomics research which may provide means to identify risk factors, treatments administered and disease outcome. This database will act as a prototype and pilot study for developing future pan African clinical databasing resources.

1.9 Thesis Overview

Chapter 1: Reviews in detail the literature underpinning research undertaken in this thesis.

Chapter 2: Details a Bioinformatics pipeline, a plethora of tools and techniques utilised to process and analyse WES data; and applies these tools to 5 family members affected by a rare familial kidney disease and one unaffected family member, leading to the high quality variant calling and genotyping. A detailed analysis of the variation identified as well functional annotation and variant prioritisation is also provided

Chapter 3: Provides comprehensive details of painstaking functional analysis methods undertaken to assess the true significance of identified variants, and implicate potential causative variant(s) and identify candidate gene (s).

Chapter 4: Unpacks the utility of clinical databasing in addressing some challenges presented by risk factors of kidney diseases in African populations.

Chapter 5: Summaries key findings of this Bioinformatics thesis and highlights potential future direction of this research.

2 Computational high throughput genomic study of rare familial kidney disease in Africa

Abstract

***Background*:** End-stage renal disease (ESRD) is a complex trait that may involve multiple processes which work together in the background of a significant genetic susceptibility. Black Africans have been shown to bear an unequal burden of this disease compared to their Caucasian counterparts. Whole exome sequencing (WES) is the application of next generation sequencing technology to determine all genetic variation in the coding regions. It is particularly effective for the unbiased discovery of highly penetrant rare variants and other functional mutations with large effect size, which are expected to explain an important fraction of the genetic etiology and pathogenesis of human disease.

***Methods*:** To elucidate the genetic factors and the mechanism underlying ESRD, WES was performed on six individuals (five cases and one control) from a large South African family of mixed ancestry with an autosomal dominant phenotype of adult-onset nephropathy characterized by early onset abnormal serum creatinine and in some cases developmental defects affecting the kidneys. Samtools, Novoalign, Picard, Genome analysis tool kit (GATK), Variant annotation analysis and selection tool (VAAST) and Ingenuity variant analysis (IVA) were applied for bioinformatics analysis.

***Results*:** From WES data of six family members; a total of 23 196 SNVs (missense, nonsense, splice site variants), 1445 insertions and 1340 deletions all of them heterozygous variants in keeping with an autosomal dominant phenotype, overlapped amongst all affected family members. Of these, only 1550 SNVs, 67 insertions and 112 deletions were present in all affected family members but absent in the unaffected family member. Further variant prioritisation based on biological parameters yielded a list of 40 variants in 35 genes, 4 novel and 6 without MAF number based on the 1000 genomes data. Copy number variants loci and Variable number tandem repeat identified did not segregate with disease in affected family members. Computational relatedness analysis revealed approximate amount of DNA shared by family members and confirmed reported relatedness. Genotyping for the Y chromosome was additionally performed to assist in sample identity.

***Conclusion*:** Next generation sequencing of five affected and one control individual from this family is sufficient to generate a list of 40 candidate etiological variants for the disease phenotype. This work clearly shows the successful application of WES for the identification of pathogenic mutations that may explain complex renal phenotypes.

2.1 Background

The precise delineation of causal variants that alter human phenotypes, predominantly diseases, is a fundamental goal of human genetics, providing crucial insights into the biology connecting genotype and phenotype and potentially facilitating the prediction of disease onset, better understanding of disease mechanism and prognosis, and new therapeutic targets. Elucidating the genetic basis of human diseases and other health-related traits has commonly relied on the oversimplified but nevertheless useful dichotomy between monogenic-simple and rare, and multigenic-complex and common diseases. Genetic variation plays a major role in both Mendelian and non-Mendialian diseases. Among approximately 2700 Mendelian diseases for which underlying genetic causes have been resolved, the overwhelming majority are caused by rare mutations that disrupt the proper function of their individual proteins. However, a substantial gap still exists in our knowledge of the relationship that exists between genotype and phenotype, and how this affects disease and other traits.

The advent of next-generation DNA sequencing has provided a means to define nearly comprehensive maps of genetic variation, including several million single nucleotide variants (SNVs), thousands of small insertion or deletion events and thousands of structural variants which are typical found in human genomes (Bentley et al., 2008; Metzker, 2010; Shendure and Ji, 2008). Most of these are common, but individual genomes also contain many thousands of rare and effectively private genetic variants (Consortium and others, 2010). Despite the existence of new methods to comprehensively catalogue human genetic variation, the identification of variants that are causal for disease or other traits remains a difficult challenge.

Although traditional gene mapping approaches such as karyotyping, linkage analysis, homozygosity mapping and gene panels have led to great insights into disease gene identification over the past few decades, they are, yet, unable to detect all forms of genomic variation (Kerem et al., 1989; Lander, 2011; Lander and Botstein, 1987). The application of these approaches is dependent on whether the disease is, for example, caused by single nucleotide mutations or by CNVs, which is difficult to predict in advance. In addition, mapping approaches often do not reduce the number of

candidate genes sufficiently for straightforward follow-up by Sanger sequencing. Also, the availability of only a small number of cases or families to study, reduced penetrance, locus heterogeneity and substantially diminished reproductive fitness are some of their drawbacks.

NGS transcends these issues and has tremendously changed the landscape of rare genetic disease research, with causative genes being identified at an accelerated rate (Kelly et al., 2015). NGS techniques have transformed genetic research, enabling rapid increases in the discovery of new functional variants in syndromic and common diseases (Gonzaga-Jauregui et al., 2012). NGS has been widely adopted by the research community and is fast being implemented clinically, motivated by recognition of its diagnostic utility and improvements in quality and speed of data attainment as well as rapidly decreasing costs (Mardis, 2011). Also, NGS methods have the potential to identify all kinds of genetic variation at a base pair resolution throughout the human genome, in a single experiment. Although WGS is still considered cost prohibitive, the application of WES using enrichment methods (Choi et al., 2009a, 2009b) has the ability to capture a highly enriched subset of the genome in which variants with large effect size, that is those that affect protein structure can be interrogated. This technique has opened cost-effective and interesting new avenues in identifying disease-associated mutations.

WES provides an unbiased and comprehensive assessment of the coding genetic variation of an individual and has been applied successfully in studies of monogenic disorders with small sample sizes (Choi et al., 2009a, 2009b). In a widely publicized case in 2011, a single exome of an affected child was used to identify a single base aetiological variant with a large phenotypic impact (Worthey et al., 2010). Also, WES has been undertaken to ascertain the genetics underlying rare complex disorders (Seidman and Seidman, 2001). Another impact of WES is to provide molecular insights into understanding the mechanisms controlling blood pressure in hypertensive patients, a common disease (Austin et al., 2012). WES has also been utilized to highlight novel insights into cancer mechanisms, by comparison of germline and somatic mutations that predispose to cancer development (Yan et al., 2011).

As highlighted, the potential of WES in identifying causative genes for rare and complex genetic disorders is enormous. In light of that, a similar approach was adopted in this thesis; WES was performed on a large South African family of mixed ancestry with a rare and difficult autosomal dominant phenotype of adult-onset nephropathy characterized by early onset abnormal serum creatinine and in some cases, developmental defects affecting the kidneys and progressing rapidly to ESRD. In this family the disease is also characterised by recurring kidney failure after transplantation. Renal biopsies performed were not very informative and failed to offer a conclusive plausible underlying cause of ESRD in this family. Therefore, WES was utilized in order to gain better insight into the underlying genetics of ESRD in this particular family. The discovery of highly penetrate rare variants and other functional mutations with large effect size, which may explain an important fraction of the genetic aetiology and pathogenesis of ESRD in this family, was sought.

Since some affected family members have progressed to ESRD and are experiencing multiple recurring kidney transplant failures, precise molecular diagnosis is clinically valuable as it may make it possible to offer genetic counseling and enable predictive testing to be offered to relatives. Perhaps it may also aid therapeutic decision-making and allow more accurate prognostic advice to be given. The utility of WES in unravelling the genetics of kidney diseases has been explored elsewhere (Choi et al., 2009a; Edwards et al., 2014; Xiu et al., 2014). Its application, however, to rare and difficulty familial kidney disease from African patients has been limited. To date and to our knowledge this is the only study using NGS techniques that has sought to explore and uncover the genetic basis of a rare ESRD and establish its pathogenesis in an African family. Because of the unexplored genetic diversity found in African patients, this may offer some novel insights into the aetiology of this kidney disease phenotype.

2.2 Materials and Methods

Ethics clearance for the study was obtained from both the University of Cape Town (HREC 521/2009) and the University of the Western Cape, and written informed consent was obtained from each participant.

2.2.1 Human Patients

Six related individuals from a South African family of mixed ancestry were enrolled in the study (Figure 2.1). The family members chosen were the ones who consented to the study, unfortunately other family members had passed away before the commencement of the study.

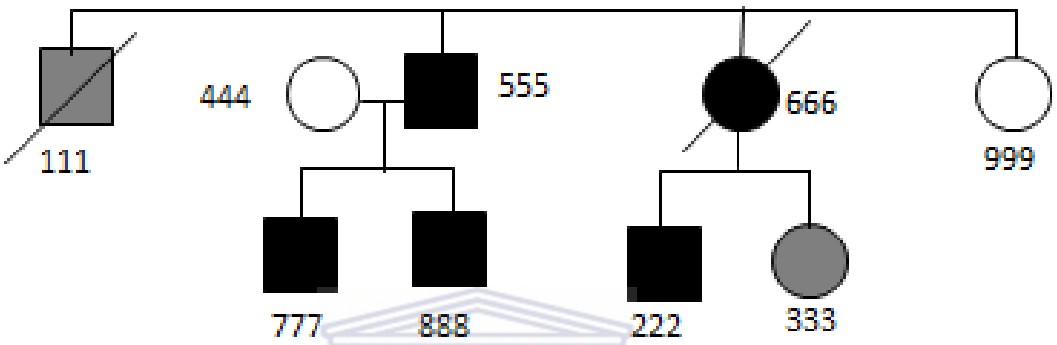


Figure 2.1 Family pedigree. Shaded black icons indicate family members affected by ESRD. Grey shaded icons indicate family members in whom disease phenotype cannot yet be determined. Crossed icons indicate family members who passed away. Unaffected people are shown by colourless circles. Squares represent male family members and circles represent female family members.

All affected family members underwent urinalysis and renal function evaluation; all of them presented with elevated serum creatinine levels as a marker of renal insufficiency however, none of them presented with any hematuria and proteinuria which was an unusual and striking finding. Two of them had kidney transplant with 2 having had at least one failed transplant. Kidney biopsy performed on one of patients showed interstitial fibrosis. After primary clinical diagnostic methods had failed to provide a clear pathogenesis of ESRD in this family an alternative approach was then sought to try and uncover the cause of the disease in this family. As a result WES was undertaken.

2.2.2 Blood collection and DNA extraction

DNA was isolated from peripheral blood lymphocytes using the salt-out method, and suspended in TE buffer. All samples underwent quality control assessment before sequencing.

2.2.3 Whole exome capture and sequencing

WES was undertaken for 6 samples obtained. Sequencing was done by a commercial company in the USA (*www.**otogenetics**.com*). DNA target enrichment was performed using Agilent SureSelect Human All Exon V4, according to the manufacture's protocol. Agilent SureSelect enrichment techniques enable the capture of genomic targets using long 120 nucleotide RNA baits which allow for efficient enrichment of regions of interest facilitating confident variant calling. Its comprehensive design targets coding regions of genes included in major databases. Enriched exome fragments were sequenced using the HiSeq 2000 platform (Illumina, San Diego, CA, USA) to get 100bp paired-end reads. Mean exome coverage of approximately 65.65× was obtained to accurately call variants at 99.41% of the targeted region. Additional quality control assessment was done on FASTQ files using FASTQC. Most of the reads had bases with Phred quality score above 20 as a result no further trimming was performed on the reads.

2.2.4 Bioinformatics analysis of whole exome sequence data

Analysis of WES data is multifaceted involving several essential steps which must be followed meticulously. In this project it was accomplished in four phases each incorporating various aspects which were carefully chosen to ensure that variant calls of highest quality were called and only good candidates were identified. These four phases are described in detail below, namely: (a) The mapping, alignment and local realignment of each exome to the human reference genome, (b) Variant and genotype calling using GATK, (c) Annotation of functional variants using VEP, ANNOVAR and (d) Probabilistic and non-probabilistic variant prioritization using VAAST and IVA respectively.

Table 2.1. Different file formats that are used in the analysis of next generation sequencing data. FASTQ file contains reads of original samples sequenced, which is usually provided by the core sequencing facility. SAM, BAM and VCF are the main file formats used for the bioinformatics analysis.

File Format	Description
FASTQ	It is a plain text format, where each single read occupies four consecutive lines: (1) The name/ID of the read, preceded by an "@"sign (2) The sequence of the read (3) A "+" sign (4) The quality scores of the bases encoded as ASCII.
SAM	Sequence Alignment/Map format. It is a TAB-delimited text format consisting of header lines which start with @ and alignment lines.
BAM	Binary Alignment/Map is a compact and indexable representation of nucleotide sequence alignments. The file is a compressed binary version of the SAM
VCF	Variant Call Format (VCF) specifies the format of a text file used for storing gene sequence variations. It contains meta-information lines, a header line, and then data lines containing information about a position in the genome. The format also has the ability to contain genotype information on samples for each position.
BED	This is a text file that contains genomic regions of interest and is useful for calculations such as coverage.

2.2.4.1 Mapping and alignment of exome reads to the human reference genome

Once generated, sequence base call files were converted to a standard file format FASTQ. FASTQ files contain millions of reads with base-associated quality scores for

each exome. The initial alignment process involves mapping reads to a best-fit location on the reference sequence. This step also associates each read with another quality score, called the mapping quality score. In this project, sequence reads were aligned to the human reference genome obtained from UCSC database (*http://genome.ucsc.edu/),* version hg19 (build 37.1), using a short read aligner program known as Novoalign (*www.novocraft.com*). Before use the human genome was indexed using Novoindex (Table 2.1). Mapped reads were stored in a sequence alignment map format (SAM) which was later converted to binary alignment map format (BAM) using SamTools (H. Li et al., 2009).

2.2.4.2 Refinement of alignments from whole exome reads

The first step to alignment refinement is removal of polymerase chain reaction (PCR) duplicates, which are reads that have the same start and end points. Duplicates arise from sequencing of identical fragments generated by PCR during library preparation. PCR errors can be introduced and propagated through unequal amplification of the library fragment template, which can lead to false positives or incorrect variant zygosity calling. In this project removal of PCR duplicates was performed using command line tools PICARD and SAMtools (Carneiro et al., 2012).

Secondly, alignment and mapping accuracy differ between algorithms. A trade-off exists between computational speed and mapping accuracy, which can lead to alignments with false positive and false negative variants. WES short reads produced by NGS instruments are difficult to map when reads contain indels, which are a significant source of false positives and false negatives variants (DePristo et al., 2011). Local realignment is performed to correct for potential alignment errors around indels. Mapping of reads around the edges of indels often results in misaligned bases creating false positive SNP calls. Local realignment algorithms use these mismatching bases to determine if a site should be realigned, and apply a computationally intensive algorithm to determine the most consistent placement of the reads with respect to the indel, whilst removing misalignment artifacts (Figure 2.2). To achieve this, the algorithm for indel realignment implemented in the GATK toolkit was used (McKenna et al., 2010) (Figure 2.2).

A third aspect to refining alignments is recalibrating base quality scores (BSQR). The raw Phred-scaled quality scores produced by base calling algorithms may not accurately reflect the true base calling error rates (Brockman et al., 2008). In such a case, the raw quality scores need to be recalibrated so that a Phred score more accurately correspond to error rates. Obtaining well calibrated quality scores is an important step, as SNP and genotype calling at a specific genomic position relies on both the base calls and the per base quality scores of the reads overlapping that position. Using SOAPsnp, per base quality scores are recalibrated by comparing a sequenced genome to the reference genome at sites with no known SNPs (R. Li et al., 2009). GATK takes into account several covariates such as machine cycle and dinucleotide context when inferring the recalibration model. Recalibrated quality scores are then estimated by adding to the raw quality scores the residual differences between empirical quality scores and the mismatch rates implied by the raw quality scores. The 1000 Genomes project adopted the same recalibrated algorithm.

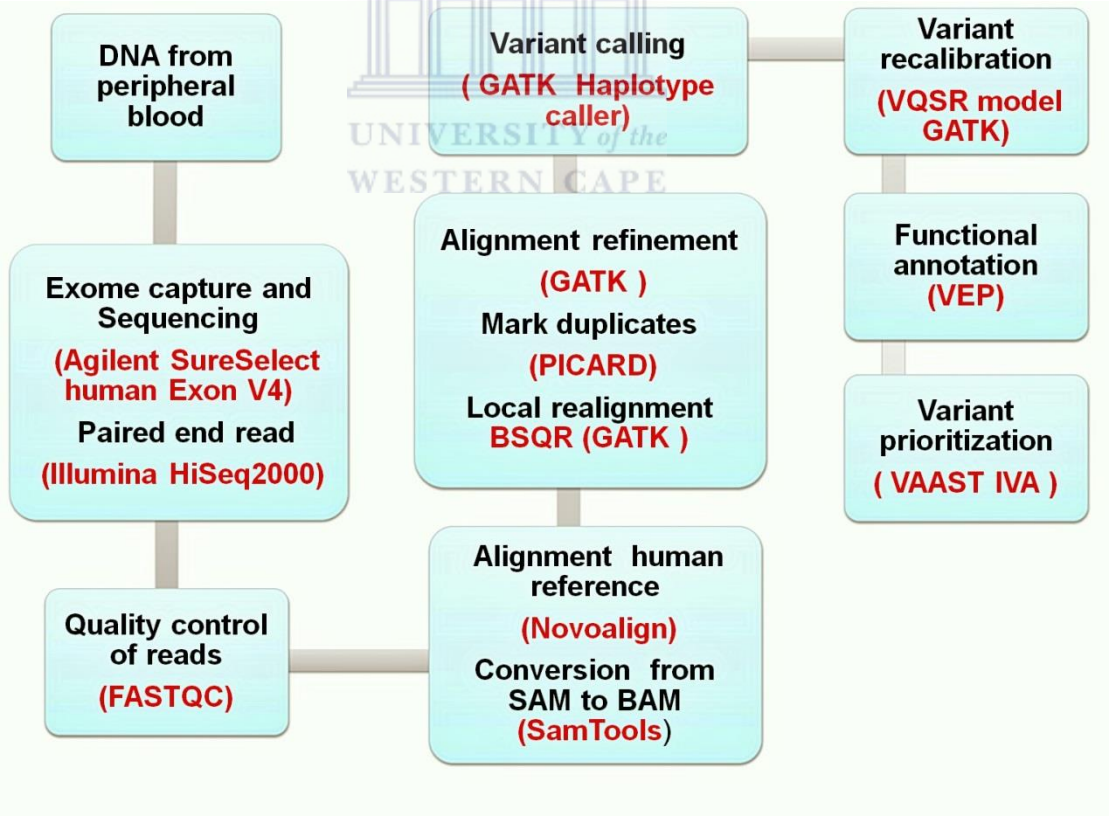


Figure 2.2 Basic workflow for WES data processing steps. The workflow shows a step-by-step breakdown of the processes involved from DNA extraction, DNA sequencing and bioinformatics analysis leading to the generation of a VCF file with ready for analysis variants.

2.2.4.3 Variant calling and statistical genotyping

Variant calling is a process of determining at which positions at least one of the bases differ from a reference sequence. In this project, multi-sample variant calling was performed using the Haplotype caller, a statistical probabilistic algorithm incorporated in the GATK (Box 2.1). Re-aligned and re-calibrated BAM files were used as input files and the output was a multiple sample variant call file (VCF). An additional strategy to reduce the number of false positive variant calls was performed using the "variant quality recalibration" in GATK. This produced a filtered and calibrated VCF that was used for variant annotation and candidate gene prioritisation.

Accessing GATK and memory assignment: Genome Analysis Tool Kit

Loading the variant caller track: Haplotype Caller

Loading the reference human genome: ucsc.hg19.fasta

Uploading all refined bam files one after the other: 999.bam……., 000.bam

Adding resources with known variants:
dbsnp_137.hg19.excluding_sites_after_129.vcf

Add Confidence call parameter 1: conf 50.0

Add Confidence call parameter 2: conf 10.0

Add a list of the targeted genomic positions: AllTracks.bed

Output variants in a VCF: output.vcf

Box 2.1: GATK's Multi sample variant calling algorithm performed in UNIX.

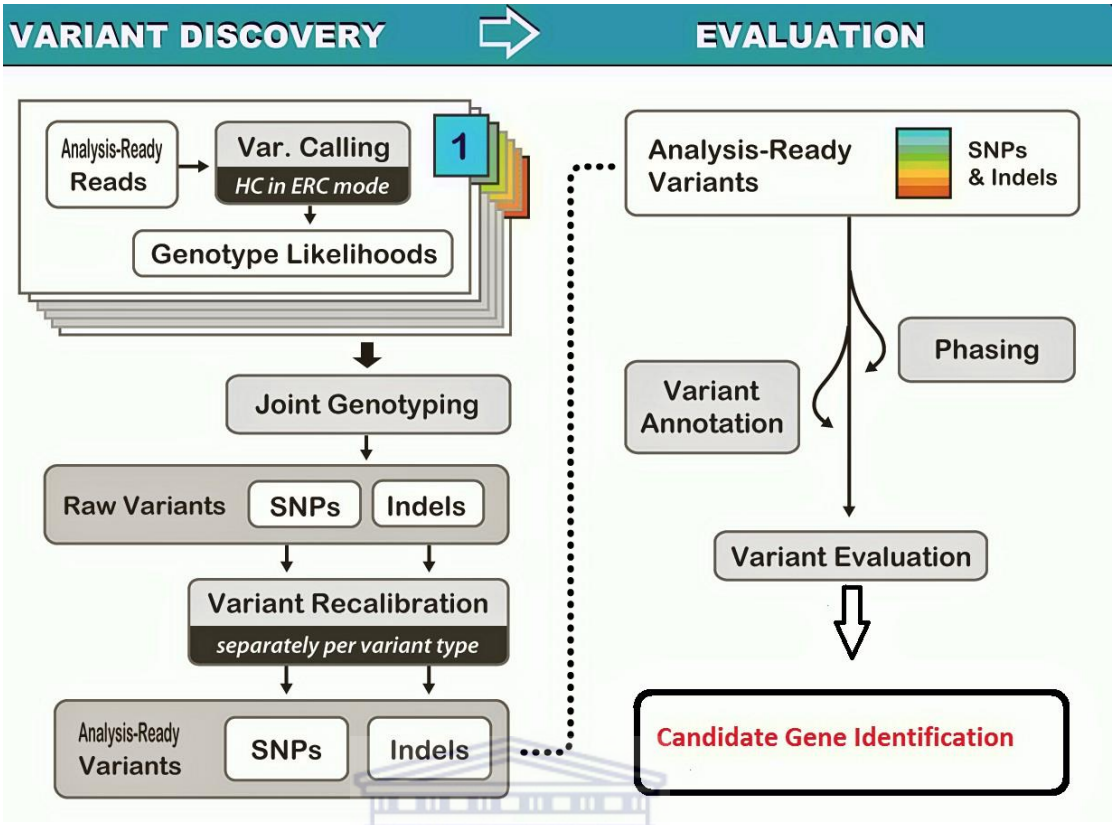


Figure 2.3 WES variant calling steps using GATK. Successive steps required to generate a list of variants from calibrated and realigned bam files.

2.2.5 Functional Annotation of identified variants

Ensembl's variant effect predictor (VEP) (Overduin, 2011) was used to annotate variants obtained from GATK (McKenna et al., 2010). The VEP determines the effect of variants on genes, transcripts, protein sequence and regulatory regions. Also, variants were prioritized further based on their genomic location; intronic, splice site, exonic as well as the type of variant, that is: SNPs, insertions, deletions, stop-loss or gain variants and whether the identified variant is known or novel using public databases namely dbSNP and 1000 genomes. ANNOVAR (Wang et al., 2010) and Seattleseq *(http://snp.gs.washington.edu)* were also used as additional annotation tools.

Table 2.2 Description of exonic variants annotations used in this project.

Annotation	Variant Explanation
frameshift insertion	An insertion of one or more nucleotides that cause frameshift changes in protein coding sequence.
frameshift deletion	A deletion of one or more nucleotides that cause frameshift changes in protein coding sequence.
frameshift block substitution	a block substitution of one or more nucleotides that cause frameshift changes in protein coding sequence
Stopgain	A nonsynonymous SNV, frameshift insertion/deletion, nonframeshift insertion/deletion or block substitution that lead to the immediate creation of stop codon at the variant site. For frameshift mutations, the creation of stop codon downstream of the variant will not be counted as stopgain.
Stoploss	a nonsynonymous SNV, frameshift insertion/deletion, nonframeshift insertion/deletion or block substitution that lead to the immediate elimination of stop codon at the variant site
nonframeshift insertion	An insertion of 3 or multiples of 3 nucleotides that do not cause frameshift changes in protein coding sequence.
nonframeshift deletion	A deletion of 3 or multiples of 3 nucleotides that do not cause frameshift changes in protein coding sequence.
nonframeshift block substitution	A block substitution of one or more nucleotides that do not cause frameshift changes in protein coding sequence.
nonsynonymous SNV	A single nucleotide change that cause an amino acid change.
synonymous SNV	A single nucleotide change that does not cause an amino acid change.
Unknown	Unknown function (due to uncertainty or errors in the gene structure definition in the database file).

2.3 Results

2.3.1 Sequencing and quality control

The average per base phred scale quality score of each sample sequenced was above 20. This is an acceptable cut-off required for analysis in subsequent steps. The reads were considered to be of high quality. No further trimming of low quality bases was required.

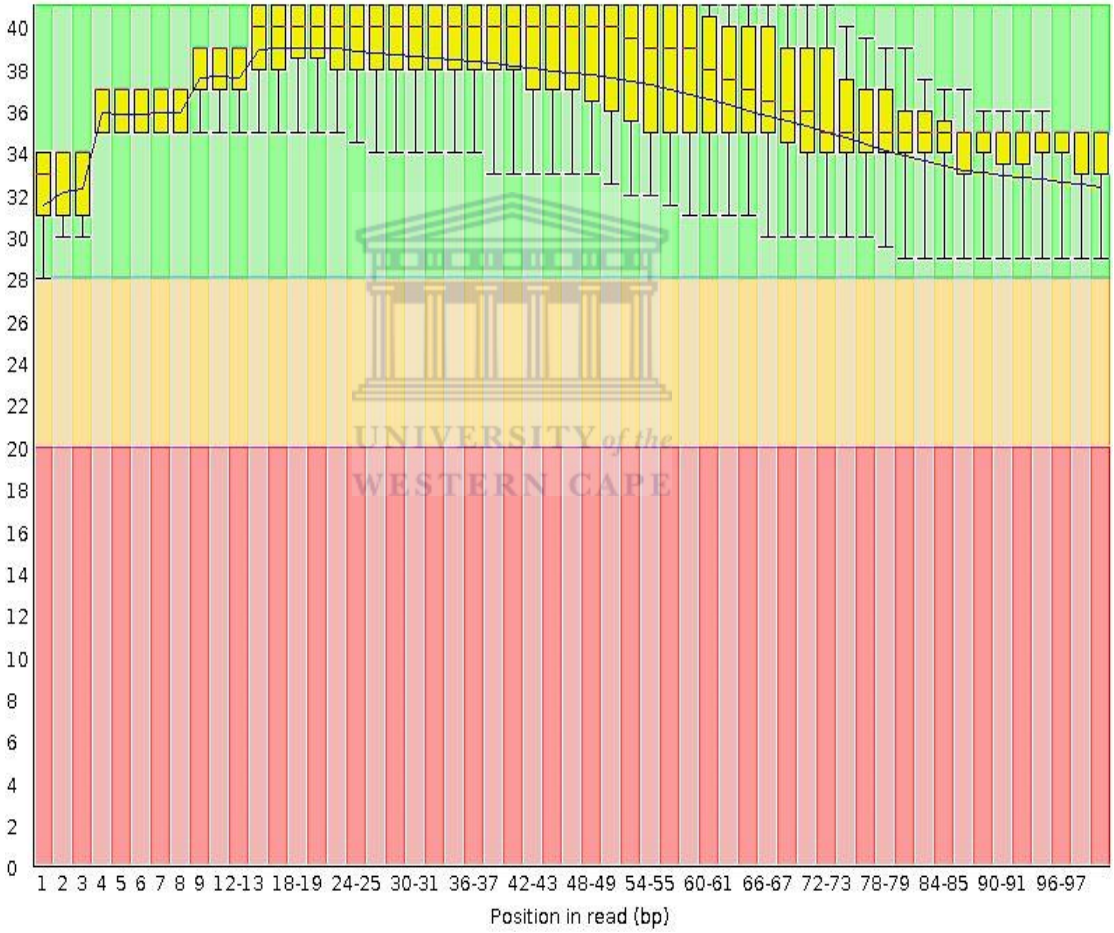


Figure 2.4 Quality control results for one of the sequenced samples (666). The quality scores show per base quality for each sequenced base position across all reads in this sample. As expected the quality of the bases was high in the middle positions of the reads and decreased towards the end as well as the beginning of the reads. Overall, per- base quality was high indicating good quality data.

Table 2.3 Summary of mapping statistics for exome sequenced samples. The mapped data is the sum of read bases that aligned to the target region. Reads that did not map to the target region were considered not mapped. Mapping and indexing of the reference genome was performed using Novoalign.

Sample	Gender	Status	Total reads	Mapped reads (%)	Mapped confidently (%)	Not mapped (%)	Mapped repetitively (%)
222	M	Affected	55 629 258	97.13	96.60	2.87	0.52
555	M	Affected	49 177 820	97.01	96.48	2.99	0.53
666	F	Affected	59 570 612	97.15	97.25	2.49	0.52
777	M	Affected	59 456 794	97.25	96.71	2.25	0.55
888	M	Affected	50 135 248	97.32	96.81	2.68	0.50
000	F	Normal	52 168 349	97.43	96.98	2.51	0.54

Targeted DNA sequencing of 6 individuals was performed; 4 affected males, 1 affected and 1 unaffected females. A total of 325 million (60 GB) high quality raw reads were obtained with an estimated average of 54 million raw reads per sample (Table 2.3). Approximately 97% of the reads mapped to the targeted regions of which ~ 96% mapped with high confidence, with a per base mismatch rate of less than 3%. There was minimal repetitive mapping averaging less than 0.5%, which may be attributed to recent methodological improvements in the hybrid capture methods and sequencing techniques. Overall, the data was of good quality.

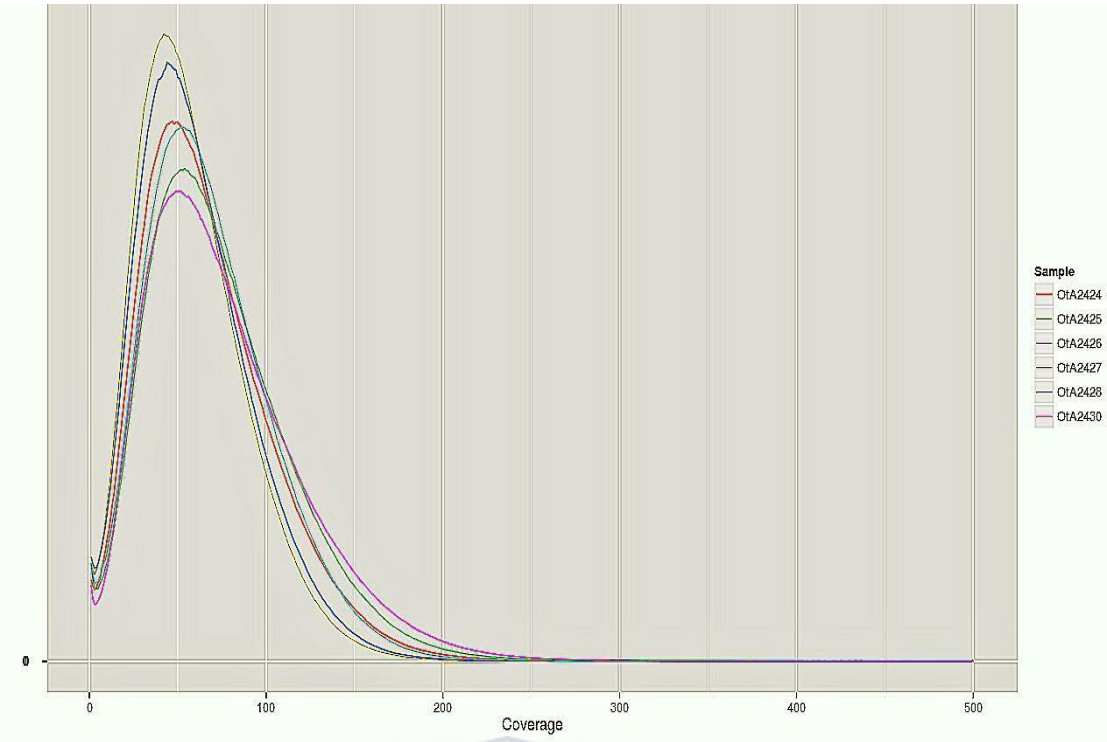


Figure 2.5 Depth of coverage distributions across the targeted region. The coverage was calculated for all mapped reads across the target region. Coverage was calculated using the Depth of Coverage tool from GATK and it was done for all sequenced exomes. The different colors show different individual exomes sequenced for which coverage was calculated.

To calculate the depth of coverage, target regions were split into 100bp non-overlapping tiles, with small tiles at target region edges. Difficult target regions were defined as those tiles with fewer than 50% of their bases covered at least 15X in the full alignments, while easy target region tiles were those with all their bases covered at least 15X in the full alignments. Overall, coverage depth for all samples averaged between 50-68× (Figure 2.5). More than 90% of targeted bases had greater than 50× coverage. This was sufficient to call SNPs and INDELs in the targeted region with high accuracy, specificity and sensitivity.

2.3.2 Distribution of variation across sequenced samples

Table 2.4 Summary of variation obtained from 6 samples sequenced. The variants were obtained from joint variant calling producing a multi-sample VCF. The multi-sample VCF file was then annotated using VEP. The list also includes intronic/UTR variants.

	Sample Name					
	222	555	666	777	888	000
Non Synonymous						
Missense	10413	9982	9972	9976	9965	10082
Stopgain	124	115	123	113	117	118
Splicing	156	147	144	155	152	133
frameshift truncation	147	148	141	162	150	158
frameshift elongation	149	136	135	137	138	146
frameshift substitution	7	7	7	6	5	6
In-frame deletion	160	145	155	158	162	153
In-frame insertion	129	121	126	118	123	126
In-frame substitution	0	1	1	1	0	0
Stoploss	35	35	29	28	31	33
frameshift duplication	9	10	9	10	10	12
In-frame duplication	3	3	3	3	4	3
Synonymous	9108	8710	8691	8688	8713	8813

Based on WES analysis of five affected and one unaffected family member, a total of approximately 100 000 variants were identified. Table 2.4 provides a summary of variation identified for each sequenced sample, broken down by the type of variation. Synonymous and non-synonymous single nucleotide substitutions were the most common with missense variants dominating averaging ~9950 per sample. Nonsense and canonical splice site single nucleotide substitutions as well as insertion/deletions were far less common amongst sequenced samples. These were composed mainly of stoploss substitution, frameshift substitution and in-frame substitutions, all combined were ~100 variants in total (Table 2.4). However, despite having smaller numbers, insertions/deletions rather than substitutions had a larger percentage of novel variants.

Since the variation in Table 2.4 also included intronic/UTRs, variants were further filtered to include only those variants in the exonic regions. As a result a total of 67 % non-synonymous SNV, 27 % synonymous SNV were obtained. Also, 6% frameshift variants were identified, 2% of them stop gain, 2 % stop loss, 1 % non-frameshift deletions and 1 % non-frame shift insertions were identified (Figure 2.6).

Interestingly, 470 exonic variants in ~400 genes were classified as variation of "unknown significance". Overall, ~14 373 were heterozygous while ~7 170 homozygous. Of the exonic variants a total of ~ 19 000 had rs SNP ids while ~1000 were novel.

Figure 2.6 Coding consequence. Missense variants were the most common type of variation identified in the targeted region, followed by synonymous mutations. Frameshift variants were also fairly frequent. Stop loss and gain variants were the other types of variants identified.

2.3.3 Functional variation shared by affected family members

Table 2.5 Variation identified in affected patients but not present in unaffected family members. Approximately 24 000 variants are shared by all affected family members and these range from missense, nonsense and splice-site variants to insertions and deletion variants.

Variation type	Shared by all affected family members	Shared by all affected family members and absent in the unaffected family member
Single nucleotide variants (missense, nonsense, splice variants)	23 196	1550
Insertions	1445	67
Deletions	1340	112

Since the inheritance model of the disease in this family is suspected to be autosomal dominant, it was imperative to look at how much variation was shared by the affected family members only, and also infer how much of this variation was not present in the unaffected family member. This was premised on the understanding that if the disease is full penetrance as we hypothesised then the causative variant can only be found in

all affected family members, and absent in the unaffected family member.

All affected family members were found to share a lot of functional variants ~ 23 196 SNVs, 1441 insertions and 1340 deletions (Table 2.5). Of the SNVs identified, ~4511 were non synonymous, 5157 synonymous, and there were 1000 splice site variants.

The variants were further analysed to identify only variants that are in the all affected family members, as already identified, but in addition are not present in the unaffected family member. ~1550 SNVs, 67 insertions and 112 deletions were thus identified (Table 2.5). The SNVs were composed of 311 non- synonymous variants, 647 synonymous variants, 73 splice variants and ~ 20 stop gain and loss variants. Overall, these were regarded as the major variants of interest.

2.3.4 Variant prioritisation using Ingenuity variant analysis

Table 2.6 Stepwise variant and gene prioritisation process. Heuristic variant filtering was performed in IVA. Common variants, predicted deleterious, genetics and biological analysis were the different filters applied to reduce the number of variants in the list to 41 in 35 genes.

Filter Type	Number of variants	Genes
Common filter (MAF<5% 1000 genomes, <5% dBsnp,<5% NHLBI)	17 189	8490
Predicted Deleterious (pathogenic, evolutionary conservation)	7 445	5077
Genetic analysis (heterozygous variants, present in **All Affected and absent unaffected**)	80	70
Biological analysis (keep only variants known or predicted to affect the disease)	**41**	**35**

In order to reduce the search space and prioritise potential causative variant(s), IVA a

non-statistical prioritisation method was used (Table 2.6). A cautious, well defined heuristic variant filtration process was undertaken. All variants shared by the affected members were compared to public databases and variants with MAF< 5% were kept for further analysis. This was based on the assertion that the causative variant is not common in the population; as such a very rare or novel variant was sought. This resulted in the number of variants reduced to 17 189 and these variants were found in 8490 genes (Table 2.6). Another filter of predicted deleteriousness was applied. This filter looks at pathogencity and evolutionary conservation to prioritise variants, further halving the number of variants to 7445 in 5077 genes. Even at this point the number of potential causative genes was still large and required further prioritisation to zero in on the probable candidate.

The genetic analysis filter was then applied. The filter prioritises variants based on zygosity and it was at this point that variants found in the unaffected family member were removed, with the remaining variation found only in affected family members. Using this filter the number of variants reduced drastically to 80 variants in 70 genes. The remaining variants were further prioritised using the biological filter resulting in 40 variants in 35 genes being prioritised (Table 2.6). These remaining variants needed to be prioritised further using a plethora of functional analysis techniques to determine the potential candidate gene.

2.3.5 Prioritised variants and their possible effects

After a comprehensive stepwise filtering using Ingenuity Variant analysis (IVA), a list of variants was generated (Table 2.7). Most of the variants obtained were single nucleotide variants, the majority of them being missense (15). Insertions (in frame (2), frameshift (2)), splice site loss (2) and synonymous (8) mutations were the other types of variants identified. The majority of variants identified had rs dbSNP ids with very low minor allele frequencies (1000 genomes minor allele frequency). About 3 novel variants were identified, these were determined based on their presence or absence from a public database dbSNP. The minor allele frequency was also considered.

Given a huge number of variants still remaining after application of the biological

filter in IVA (Table 2.7), and the importance attached to the results as they would be required for application in clinical settings, additional prioritisation methods needed to be sought in order to reduce the list to manageable set supported by a wide pool of biological evidence. Variant prioritisation and candidate gene identification is not a trivial task. The process needs to be undertaken meticulously. Therefore, a range of functional analysis and variant prioritisation tools were applied and this work is explained in more detail in chapter 4.

Table 2.7 A list of prioritised genes from IVA analysis. The variants were prioritised following a stepwise filtering process that involved the removal of common variants, retaining those predicted to be deleterious and also predicted to be involved in kidney disease.

Chr	Chr position	Gene	Variation type	Translation impact	Protein variant	dbSNP rs Id	MAF
1	32164127	COL16A1	SNV	Missense	p.T116M	rs34091659	0.0056
1	32204991	BAI2	SNV	Missense	p.P805T	Novel	0
1	32670637	CCDC28B	SNV	Missense	p.G2822G	rs7543181	0.0116
1	34015796	CSMD2	SNV	Missense	p.G2966G	rs61734268	0.0194
1	35575933	ZMYM1	SNV	Synonymous	p.P282P	rs374134267	0
1	35863125	ZMYM4	Insertion	in-frame	p.K1060	Novel	0
1	43919081	HYI	SNV	Splice Site Loss	p.Y96C	rs14236920	0.00082
1	100347219	AGL	SNV	Missense	p.S743S	rs373513564	0.0012
1	109325118	STXBP3	SNV	Missense	p.R295Q	rs2275344	0.0184
1	153748132	SLC27A3	SNV	Synonymous	p.A100A	rs373105501	0.0022
1	154842199	KCNN3	Insertion	in-frame	p.Q77_Q80dup	Novel	0
2	37450350	CEBPZ	SNV	Missense	p.A615V	rs34983085	0.0018
5	140773738	PCDHGB1	SNV	Missense	p.P453L	rs115102808	0.0036
6	151670287	AKAP12	SNV	Missense	p.P254L	rs73780648	0.0172
6	151917596	CCDC170	SNV	Missense	p.H532Y	rs201625561	0.0007
6	152469200	SYNE1	SNV	Missense	R8319Q	rs148008634	0.0006
6	152749340	SYNE1	SNV	Splice loss	p.R1666K	rs111428582	0
11	131240728	NTM	SNV	Synonymous	p.H9H	Novel	0
13	35923326	NBEA	SNV	Synonymous	p.E1995E	rs151318906	0.0034
13	110845216	COL4A1	SNV	Missense	p.R476W	rs369960952	0.0002
16	75269325	BCAR1	SNV	Missense	p.R281H	rs16957558	0.296
16	77246091	SYCE1L	Insertion	Frameshift	p.E164fs	rs371551639	0.0018
19	4311942	FSD1	SNV	Synonymous	p.Y198Y	rs34953789	0.0092
19	6222552	MLLT1	SNV	Synonymous	p.S230S	rs139655596	0.0128
19	8145928	FBN3	SNV	Missense	p.R2471H	rs3848570	0.0222
19	10395208	ICAM1	SNV	Missense	p.P352L	rs1801714	0.00072

2.3.6 Structural variation inference from exome reads

WES data was also used to investigate co-segregation of copy number duplication or deletion using copy number inference from exome reads (CONIFER) (Krumm et al., 2012). Table 2.8 shows CNVs identified although none of the copy number variants co-segregated with affected members. Following the work of (Kirby et al., 2013), who investigated the genetics of a rare kidney disease MCKD1, a condition that rapidly progresses to ESRD ultimately requiring renal transplantation for affected patients. After painstaking work the researchers identified and implicated variable number tandem repeats (VNTRs) in the MUC1 gene as causative for the disease. A similar approach was adopted in this study and the entire targeted region was scanned. Any tandem repeats that might segregate with the disease in this family were sought. A short tandem repeat profiler for personal genomes (lobSTR) was utilized. None of the identified tandem repeats segregated with affected family members.

Despite the drawbacks of computational tools in identifying structural variation from WES data, the meticulous way that it was undertaken in this project led to the conclusion that although structural variation maybe important in elucidating the genetics of rare diseases, in these particular patients it may not be the case. This exploration of CNV has added weight to our hypothesis that a rare or novel fully penetrant SNV or small INDEL most likely to be implicated as disease-causing in this family.

Table 2.8 Copy number variants detected in sequenced samples. CNVs identified were computationally inferred. None segregated with the disease in affected individuals.

Sample Id	Chromosome	Chromosome position	CNV type
222	4	69342036-69417736	Deletion
222	16	3705885-3712955	Deletion
222	16	2223808-2224650	Duplication
666	11	8706408-8715717	Duplication
777	8	39311494-39332289	Deletion

2.3.7 Relatedness analysis using Whole exome data

Table 2.9 Amount of shared DNA amongst family members. The analysis was performed using Plink a tool widely using for Genome wide analysis (GWAS). A multi sample VCF file obtained from variant calling was used as in put for the analysis.

Sample ids	Relationship	Amount of DNA shared
555 – 777	Parent/child	0.5000
555 – 888	Parent/child	0.5000
777 – 888	Siblings	0.5290
666 – 222	Parent/child	0.5000
555 – 999	Siblings	0.5103
666 – 999	Siblings	0.5000
999– 777	Aunt/Nephew	0.3331
999 – 888	Aunt/Nephew	0.3119
999 – 222	Aunt/Nephew	0.3209
555 – 222	Aunt/Nephew	0.2194
666 – 777	Aunt/Nephew	0.3081
666 – 888	Aunt/Nephew	0.2835
777 – 222	Cousin	0.2390
222 – 888	Cousin	0.2200
555 – 666	Siblings	0.5000

In order to assess the amount of DNA shared by family members a multi sample VCF file obtained from GATK's haplotype caller was used as an input to Plink (*http://pngu.mgh.harvard.edu/purcell/plink/*), which is a widely used tool for GWAS data analysis (Purcell et al., 2007). The father 555 shared approximately 50% of DNA with his sons 777 and 888 (Table 2.9). Similarly, 666 shared approximately 50% of DNA with her son 222. Overall, brothers and sisters shared approximately 50% of their DNA as is expected and first generation cousins shared approximately 30% DNA. Relatedness analysis is important to perform since an inherited genetic disorder was being studied: absolute certainty is required to establish true paternity, and to confirm the relatedness self-reported by family members. This is important when it comes to segregation analysis and validation of candidate variants. In the family under

study reported relatedness was confirmed by genetic relatedness.

2.4 Discussion

The cost of sequencing DNA has declined steeply since the advent of next-generation short read technologies (Wetterstrand, 2013). It is now at the point where realistically large cohorts of whole human genomes can be sequenced for further analysis. Currently, investigations of disease-causing variation continue to focus on the protein-coding exome, which is a small fraction of the whole genome (Meynert et al., 2014). It contains fewer repetitive elements than non-coding regions and contains most of the causal disease variants identified to date (Goldstein et al., 2001; Ng et al., 2008). Additionally, experimental approaches to determine the function of candidate disease variants at protein coding or transcript splice sites are well developed, recognized and accepted by the research community (Meynert et al., 2014). For these reasons, exome centric analysis will remain common in research and is increasingly used in clinical genetic settings (Yang et al., 2013).

The targeted capture followed by sequencing of specific regions, such as the human exome (WES), has proven to be a cost-effective and productive strategy for the identification of single nucleotide polymorphisms (SNPs), small insertions and deletions in this rich vein of the genome. Genetic variation in protein-coding portion of the genome is of significant interest in the study of human health. The focus on coding exons is due largely to the common credence that the exome harbors the most functional variation (Botstein and Risch, 2003; Stenson et al., 2014). This is based on the observation that mutations that cause Mendelian diseases occur primarily in genes and result in altered protein function (Botstein and Risch, 2003; Stenson et al., 2014). Mutations that cause amino acid substitutions, including changes to nonsense codons, in their respective genes are the most frequent type of disease mutation (Botstein and Risch, 2003; Stenson et al., 2014). In addition, small indels in genes account for almost a quarter of the mutations in rare diseases identified to date (Botstein and Risch, 2003; Stenson et al., 2014). Our understanding of functional variation is an important step towards an era of precision medicine, where a physician can inform patients of their disease susceptibilities or aetiology based on their genome sequences. Consequently, if the exome harbors much of the functional variation responsible for a

person's phenotype, then identification and characterization of the individual's variation in the exome could enable individualized genomics and personalized medicine.

Clinical whole exome sequencing (CWES) is rapidly becoming an essential part of the clinical approach for individuals with rare diseases, and is being applied to a wide range of clinical presentations that require a broad search for causal variants across the spectrum of genetically heterogeneous disorders (Lee et al., 2014). However, a major challenge of undertaking CES is the interpretation of the variants in the context of the phenotypic data provided. Currently, available options for genetic testing include Sanger sequencing of candidate gene(s), next generation sequencing of a selected panel of candidate genes. Application of WES, however, yields data on a far greater number of genes, including those not initially considered candidates for the phenotype being investigated. For instance, this allows a diagnosis to be made where detailed phenotype data such as a kidney biopsy are lacking or are inconclusive, but variants may be detected that predict disease pathogenesis.

Thus, in patients who are suspected of having an inherited renal disease, genetic investigations may provide a more definitive diagnosis than renal biopsies (Adam et al., 2013). Previously, genetic diagnosis in kidney disease has been limited to patients and families, where clinical or histological data display distinctive features that suggest one or a small number of candidate genes that can be sequenced individually. However, with the advent of massively parallel next-generation sequencing, a large number of genes can now be investigated in a single patient and/ family at a cost that can reasonably be covered by healthcare providers. The benefit of this approach is that it allows a precise molecular diagnosis to be made even in patients where clinical data are lacking or non-specific, and can sometimes avert the need for invasive tests such as a kidney biopsy.

In this study, WES was performed using NGS techniques to interrogate approximately 1% of the entire human genome, the exome. The overarching aim was to identify rare or novel functional variation that could assist in unraveling and elucidating the genetic basis, and help explain disease pathogenesis in a rare familial clustered kidney disease. The disease in this family is characterized by a rapid progression to ESRD,

which ultimately requires lifesaving RRT. However, despite accessing RRT, the affected family members who have undergone kidney transplantation experience recurring kidney transplant failures. After a thorough, well-designed bioinformatics analysis process undertaken to call variants in all the samples sequenced it was found that, interestingly, all the affected family members shared a glut of functional variation, in excess of 20 000 variants (Table 2.5). Even though common variation is expected by virtue of relatedness in the amount of DNA shared (Table 2.9), however, some variation shared may co-segregate with the disease and these variants form the set of candidate aetiological variants for the disease.

Careful, painstaking filtering of these variants using IVA and different filtering strategies, including removal of variants that are shared by both affected and unaffected (Table 2.6) and biological parameters, yielded a reduced list, with 41 variants in 35 genes. Of interest, we observed multiple variants in some single genes for al1 affected family members, which could mean that in each affected individual several different variants may possibly work in combination to cause the phenotype, a potential scenario that requires careful attention.

Utilization of various tools to identify structural variation that co-segregates with affected status in this family showed that none of the copy number duplication/ deletions and tandem repeats that were identified segregated with disease status. Thus, while WES may provide a method to identify CNV and tandem repeats regions, determining these regions with high sensitivity, specificity and accuracy can prove difficult (Krumm et al., 2012; Tan et al., 2014; Yoon et al., 2009). The reasons and remedies for this are currently an active area of research (Robinson et al., 2011) and are beyond the scope of this thesis.

The list of genes in Table 2.7 contains genes such as COLA4, COLA16, FBN3 and ICAM1, which have been implicated in rare kidney diseases (Genovese et al., 2014; Lin et al., 2014; Malone et al., 2014; Xiu et al., 2014). In WES analysis because a large number of genes are tested simultaneously, the prior probability that any one of the genes or variants identified is responsible for disease is low. While variants known to be common in the healthy population can usually be excluded from consideration, it is not always possible to determine whether a rare or novel variant identified in a

patient is actually pathogenic. This is because, even where variants lead to a change in the amino acid sequence of a protein, there is no guarantee that the change in protein function will eventually cause a disease. Typically, several lines of evidence including linkage data, co-segregation analysis, comparison with mutation databases and functional and structural data about the protein are needed to make a diagnosis confidently (Park et al., 2013). Over time, as more extensive sequence databases, incorporating data from larger numbers of patients and healthy individuals, become available to researchers and clinicians, the power of WES approaches is likely to increase.

Results obtained in this study highlight the need to have better means to define the significance of variants obtained. For instance, about 450 variants of unknown significance were identified in approximately 300 genes. While this cannot entirely rule out the possibility of pathogenic involvement of these variants, only "functional studies" can truly assess the significance of these variants. This is also complicated further by the fact that all affected family members have plausible disease-causing variants in multiple genes (Table 2.7). In other words, detailed functional analysis process to assess the significance and identify highly plausible disease causative variants is warranted and this is undertaken in Chapter 3.

2.5 Conclusion

Based on our findings, several conclusions can be drawn: firstly, WES data analysis offers a viable, noninvasive approach that can be utilized to expedite the discovery of disease causing variants that may be used for molecular diagnosis of rare inherited genetic diseases. Secondly, in our study, WES of 1 unaffected and 5 affected family members, combined with a stepwise variant filtering strategy, led us to prioritize 41 variants in 35 genes. To reduce the multitude of variants generated by WES in-order to identify plausible causative variants additional functional analysis is required.

3 Functional analysis and candidate gene prioritization

Abstract

***Background*:** Pinpointing precisely the genomic variation in human genomes that causes disease in the era of next generation high-throughput sequencing is a major challenge of genetic diagnosis. Refining our ability to interpret variation responsible for Mendelian and complex disorders provides an opportunity to resolve elusive genetic disorders. Without rigorous evaluation of potential causative variation, the number of false positive reports of causality may increase. This would hamper the translation of genomic research findings into the clinical diagnostic setting. Therefore, a multi-factorial approach is required to integrate gene and variant level information to support any proposed causality.

***Methods*:** After exome sequencing was performed and a list of variants produced, variants were assessed further with respect to their plausible involvement in ESRD. A combination of tools was used to identify potential causative variant(s) and candidate gene(s). A heuristic step-wise analysis in Ingenuity variant analysis (IVA) was utilised, this filtered variants based on careful chosen predefined filters. A statistical probabilistic framework implemented in Variant Annotation Analysis and Selection Tool (VAAST) was also used and its results were compared to those of IVA. VarElec was utilised to infer if potential candidate genes had direct or in direct links to ESRD. Ingenuity Pathway Analysis (IPA) was undertaken to infer functional and toxic pathways that are significant. Protein-protein interaction networks were inferred using STRING and evolutionary conservation analysis was undertaken using the UCSC browser. Potential causative variants were visualised using integrated genomic viewer IGV.

***Results*:** From a total of more than 85 000 variants, I prioritised 3 novel indels and 16 missense variants in 10 genes (FBXL21, SYCE1L, KCNN3, COL4A1, ICAM1, COL16A1, ZMYM1, STXBP3, ANXA9, and CEBPZ). These were identified in all affected family members and were consistent with autosomal dominant inheritance. Of these, only 3 very rare heterozygous missense variants in 3 genes COL4A1 [p.R476W], ICAM1 [p.P352L], COL16A1 [p.T116M] were considered potentially disease causing. These variants segregated with the disease in all affected family members and none of them were detected in the unaffected family member.

***Conclusion*:** The findings show a successful application of WES with extensive variant filtering for the identification of plausible pathogenic mutations, illustrating the power of molecular genetic diagnostics techniques that may explain complex renal phenotypes.

3.1 Background

With the costs of sequencing plummeting, whole exome sequencing is being widely used in the identification of pathogenic variants for Mendelian diseases and the discovery of susceptibility loci for complex diseases (Girard et al., 2011). High-throughput exome sequencing can generate detailed catalogues of genetic variation in both disease patients and the population in general (Sadee et al., 2014). However, for this technique to have a huge medical impact, candidate disease-causing or disease-associated genetic variants must be reliably delineated from the broader background of variants present in all human genomes that are rare, potentially functional, but may not actually be pathogenic for the disease or phenotype under investigation (Cooper and Shendure, 2011; Sadee et al., 2014). Identification and accurate assessment of disease-causing variants from a long list of candidates is crucial for clinical applications such as diagnosis, newborn screening, carrier screening, selection for mutation-specific therapy, and association between genes and rare familial diseases (Bamshad et al., 2012; Ng et al., 2008, 2009b).

Pinpointing genetic variants underlying human inherited diseases is the primary step towards understanding the pathogenesis of human diseases (Cooper and Shendure, 2011). The majority of genetic variants captured by exome sequencing are non-synonymous, single nucleotide variants (SNVs) whose effect may change protein-coding sequences, thereby affecting their function and plausibly causing diseases (Bamshad et al., 2012). Unfortunately, our ability to interpret the impact of individual variants on diseases phenotype has not kept pace with the ease with which they are identified (Bell et al., 2011). It has been demonstrated that among the large number of variants obtained, the alteration of the function of a gene hosting a variant does not necessarily mean that the variant is pathogenic for the disease being investigated (Bodmer and Bonilla, 2008; Wu and Jiang, 2013).

An analysis of 406 published severe disease mutations observed in 104 newly sequenced individuals reported that 27% of these were either common polymorphisms or lacked direct evidence for pathogenicity (Bell et al., 2011). Other studies have identified numerous alleged disease-causing variants in the genomes of population

controls (Norton et al., 2012; Xue et al., 2012). In other cases, follow-up studies of high profile reported mutations have cast serious doubts on initial reports assigning disease causality to sequence variants (Hunt et al., 2012; Weng et al., 2005). As the volume of patient sequencing data increases it is critical that candidate variants are subjected to rigorous evaluation to prevent further miss-annotation of their pathogenicity.

Several lines of evidence are required to implicate a variant as disease causing (Taylor et al., 2015). Evidence implicating a variant as disease causing must be assessed thoroughly. For instance, healthy individuals carry many rare protein disrupting variants, and about half carry at least one de-novo protein-altering mutation (Hansen et al., 2015; Tennessen et al., 2012). Such variants are, therefore, not typically sufficient proof of causality when observed in a disease case, even if present in well-established disease genes (Cooper and Shendure, 2011; Gayà-Vidal and Albà, 2014). With both established and newly implicated disease genes, researchers should evaluate the statistical support for association. Also, in family-based studies co-segregation of candidate variants with disease status must be evaluated (Gayà-Vidal and Albà, 2014). Given that a separate, unobserved pathogenic mutation may lie on the same haplotype as the candidate variant, segregation analysis alone cannot definitively implicate a specific variant as pathogenic, but at least under an assumption of complete penetrance, lack of segregation can exclude non-pathogenic variants from further consideration (Gayà-Vidal and Albà, 2014). Measures of evolutionary sequence conservation have demonstrated value in prioritizing candidate variants and are utilized widely as indicators of deleteriousness for both protein-coding and non-coding variation (Cooper and Shendure, 2011). Some classes of variation, such as truncating or splice-site disrupting variants are more likely to be damaging than others, such variants are also enriched for sequencing and annotation errors and need to be thoroughly interrogated prior to assigning pathogenicity (MacArthur et al., 2012).

Therefore, extra care should be taken when returning genetic results back to the participants, and consider the serious implications of naming a variant as disease-causing to a family who may want to use the information for diagnostic and prognostic purposes; and even for some severe diseases in prenatal diagnostics which

could affect decisions of parents to terminate pregnancies. It is therefore from an ethical perspective an extremely large responsibility to do as much as possible to ensure that the identified variants are highly likely to be related to the disease.

Hence, in order to access whether a variant is causative for a disease, it is not enough to only predict the functional damaging effects of the variant. Detailed evaluation of evidence for variant implication should also focus on the statistical evidence from both genetic and functional analysis. Furthermore, a combined assessment of the genetic, experimental and informatics support for individual candidate variants should be performed. Such assessments should be performed even if the genes or variants have been previously reported. Prior evidence should be continuously re-evaluated with newly available information. As highlighted above, assigning pathogenicity and implicating a variant as truly causal for a disease is a complex, multifaceted undertaking. The previous Chapter provided a detailed step by step analysis of the exome sequencing data leading to a list of variants. This Chapter provides a detailed description of a combinatorial analysis utilized to evaluate the evidence supporting confident identification of potential disease causing variants(s) from a list of prioritized variants detected in this family.

3.2 Methods

Once a list of variants has been generated, as described in Chapter 2, a process of identifying candidate gene(s) begins. In this project, given the difficult diagnosis and rarity of the disease whose genetic basis was being investigated, a combination of tools was used. Lack of clarity in the terms used to define types of sequence variants is another cradle of contention in human genetics. In Table 3.2 is a description of the terminology followed in assigning pathogenicity in this project.

Guidelines for establishing the significance of variation exists and these can be separated into several disparate categories (Table 3.2).

Table 3.1 General steps that may be followed for implicating sequence variants in human disease. These guidelines were followed in this study to identify potential causative variant(s) (MacArthur et al., 2014).

General guidelines	Assessment of evidence for candidate disease genes	Assessment of evidence for candidate pathogenic variants
Describe and assess clearly the available evidence supporting prior reports of a gene or variant implication.	Evaluate genes previously implicated in similar phenotypes before exploring potential new genes.	Evaluate if a variant is co-inherited with disease status within affected families.
	Investigate gene products which interact with proteins previously implicated in the disease of interest.	Avoid assuming that implicated variants are completely explanatory in any specific disease case.
	Investigate if the gene and/or gene product function is demonstrably altered in patients carrying candidate mutations.	Recognize that strong evidence that a variant is deleterious is not sufficient to implicate a variant as causal for a disease.
Take advantage of public data sets of genomic variation, functional genomic data and model-organism phenotypes.	Report a new gene as confidently implicated only when variants in the same gene and similar clinical presentations have been confidently implicated in multiple unrelated individuals	Predict variant deleteriousness with comparative genomics approaches, but avoid considering any single method as definitive.
	Where possible use non-human animal models with a similarly disrupted copy of the affected gene and see if a phenotype consistent with human disease state is seen.	Use multiple methods as independent lines of evidence for implication of a potential causative variant.
		Check if the variant is found at low frequency and consistent with the proposed inheritance model.
		Check if the variant is found at the location within the protein predicted to cause functional disruption.

Guidelines for establishing the significance of variation exists and these can be separated into several disparate categories (Table 3.2).

Table 3.2 Terms used to describe DNA sequence variation.

Term	Description
Pathogenic	A variant that contributes to a disease, but is not necessarily fully penetrant (i.e., may not be sufficient in isolation to cause disease).
Implicated	A variant that possesses evidence consistent with a pathogenic role with a defined level of confidence.
Associated	Variant that is significantly enriched in disease cases compared to matched control cases.
Damaging	A variant that alters the normal levels or biochemical function of a gene or a gene product.
Deleterious	A variant that reduces the reproductive fitness of carriers.

3.2.1 Statistical probabilistic variant prioritization

While heuristic, stepwise filtering has proven successful in identifying candidate causative genes in a number of disorders (Choi et al., 2009a; Kumar et al., 2011) these methods are limited in that they do not provide any measure of statistical uncertainty for a given candidate variant identified. Using custom scripts, I applied VAAST, a new tool that uses multi-parameter likelihood equations to compare allele frequencies between affected and unaffected in combination with modelling variant severity by an amino acid substitution analysis to provide a top hit list of variants (Yandell et al., 2011). Each variant will have an associated VAAST ranking score and a P-value. The P-value is a measure of the probability that a variant is statistically significant in affected compared to the unaffected individuals.

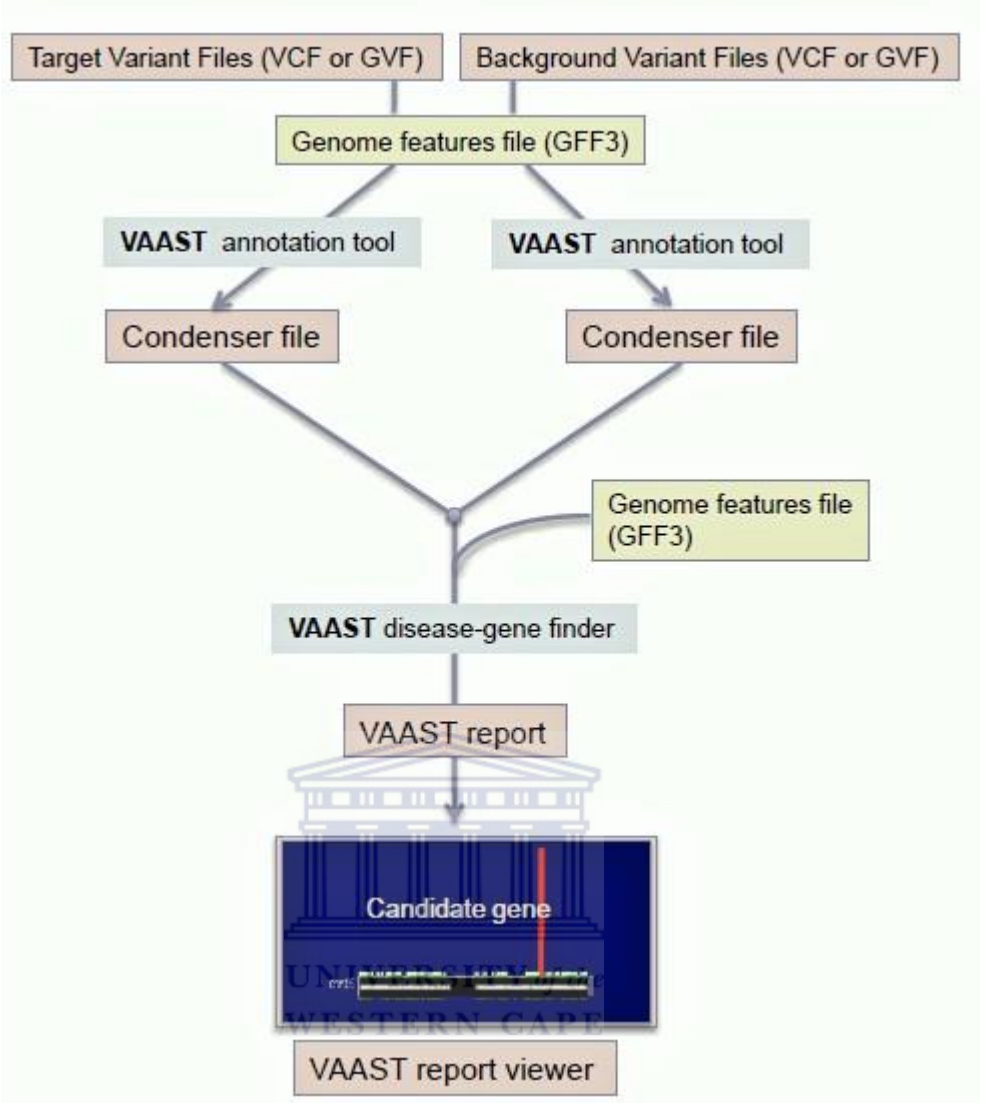


Figure 3.1 VAAST search steps followed to identify potential candidate genes. A multi-sample VCF was provided as input. An in-house script was utilised to separate variants identified in affected only and absent in the normal sample as well as implementing the probabilistic model used.

3.2.2 Ingenuity Variant Analysis

IVA was also used to identify causative variants from WES data (Ingenuity Variant analysis ™ software (www. qiagen.com/ingenuity)). IVA annotates and interactively filters data using several filters such as biological context, statistical association, genetic analysis, common and high confidence variants (Figure 3.2). Using the confidence filter IVA excludes all variants that do not pass a particular threshold, for instance a Fred quality score of 20. Thus, all variants whose quality score is below 20 are removed from further analysis. Secondly, the common variant filter looks at MAF:

since rare variants were sought, a MAF cut-off of more than 5% was used to exclude common variants. The predicted deleterious filter includes only those variants that are predicted to be pathogenic. The filter also takes evolutionary conservation into account. The genetic filter uses zygosity to exclude variants. For instance, if a disease is hypothesised to follow an autosomal dominant inheritance pattern, homozygous variants are excluded from further analysis as a heterozygous variant maybe sought. Finally, the biological filter keeps only those variants that are known and/or predicted to be involved in the disease of interest.

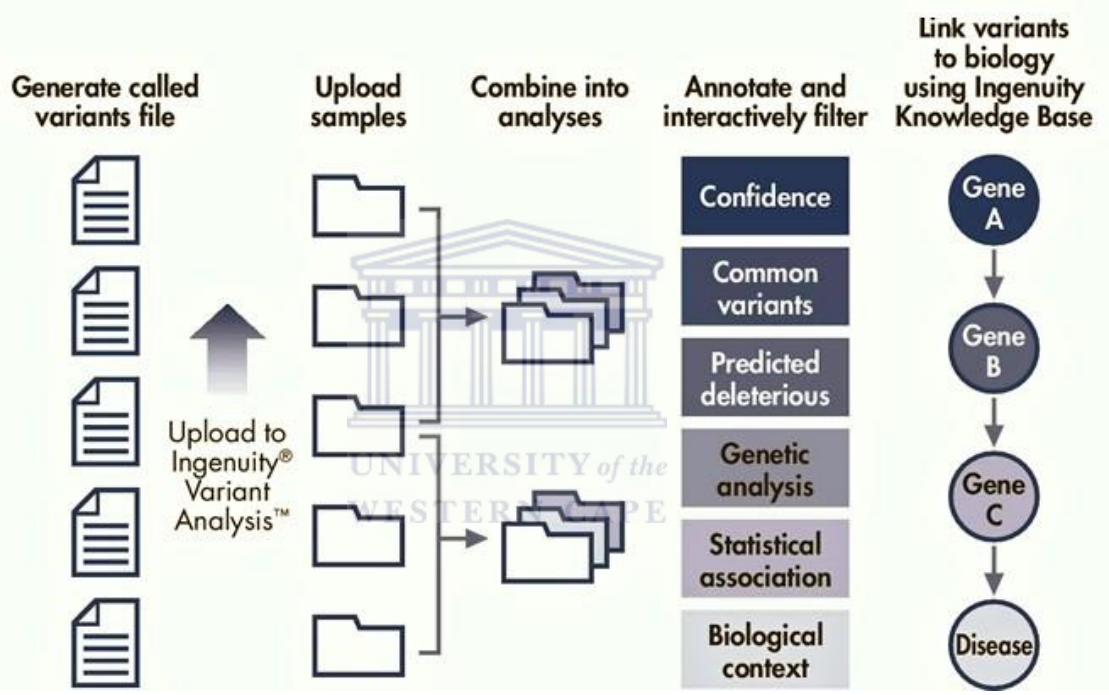


Figure 3.2. Workflow showing steps followed in candidate gene identification using IVA. VCFs files for each sample as well as a pedigree file were used as input. The variants were filtered in such a way that only variants present in affected family members only and absent in the unaffected family were prioritised (Ingenuity Variant analysis ™ software (www. qiagen.com/ingenuity).

3.2.3 VarElect

VarElect is a rapid prioritization of variant genes based on disease/phenotype of interest (Belinky et al., 2015). It provides a robust algorithm for ranking genes and predicting their likelihood to be related to a disease. To identify potential disease-causing genes the algorithm leverages the rich information within a leading human gene database (GeneCards, *http://www.genecards.org/*), the human disease database

(MalaCards, http://www.malacards.org/) and the unified human biological pathways database (PathCards, http://pathcards.genecards.org). Using VarElect, an input list of genes with variants can be narrowed down to the top 1-10 genes that are potentially associated with a particular disease. The algorithm acts jointly on the gene list and phenotype/disease keywords, and produces a list of prioritized, scored, and contextually annotated genes, whilst providing direct links to supporting evidence and further information. The degree of mutual linking is quantified via endogenous search scores.

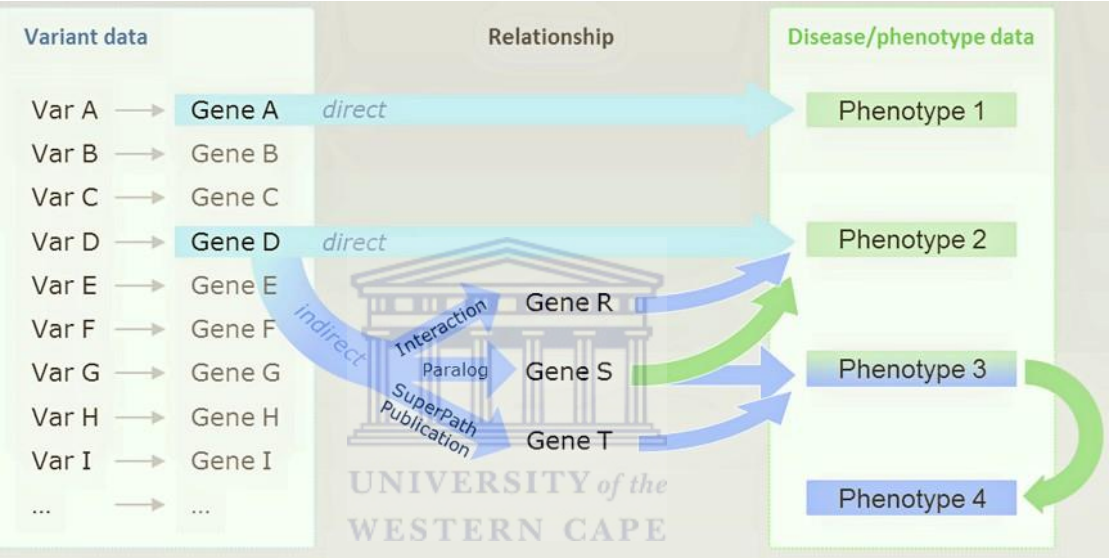


Figure 3.3 Steps followed in candidate gene prioritisation steps using VarElect. A list of genes with potential causative variants is provided as an input into VarElect (Belinky et al., 2015). The variants are inferred for their plausible direct involvement to kidney disease. A list of genes showing their evidence to disease phenotype and a classifying score showing the strength of this relationship is provided.

3.2.4 Pathway Analysis (IPA)

IPA is an application used for analysis, integration, and interpretation of data obtained from experiments, such as NGS. The analysis and search tools uncover the significance of variants and identify new targets or candidate genes within the context of biological systems. Powerful algorithms are utilized to predict downstream effects on biological and disease processes, identify regulators, mechanisms, functions, and pathways relevant to analyzed genes (IPA® QIAGEN Redwood City, www.qiagen.com/ingenuity).

3.2.5 Protein-protein interaction and other networks (STRING)

STRING is a database of known and predicted protein interactions. Protein-protein interaction networks are useful for the system-level, or mechanistic understanding of cellular processes. Such networks can be used for filtering and evaluating functional genomics data and for providing an intuitive platform for annotating structural, functional and evolutionary properties of proteins. The interactions include direct physical and indirect functional associations. They are derived from four sources, genomic context, high throughput experiments, conserved experiments and previous knowledge (*http://string-db.org/*).

3.3 Results

3.3.1 Beyond "the one hit theory"

Table 3.3 shows a striking pattern of variation that was observed in this family. As hypothesised and also based on the rarity of the phenotype, a high impact variant probably in a single gene was sought. But a number of variants located closely together on the same chromosome were identified. Four genes; COL16A1, BAI2, ZMYM1 and ZMYM4 were positioned adjacent to each other on chromosome 1 (Table 3.3). Two of these variants are novel, another had no MAF reported and the remaining two are very rare variants (Table 3.3). Considering the variant filtering strategies utilized in this project, these variants were only identified in the affected family members and absent in the unaffected. Given the close proximity of the genes the probability of recombination occurring across different generations is very low. Thus, a haplotype rather than a single variant could be inferred to be potentially segregating with disease status in this family. Also, two different variants were identified in the gene SYNE1 (Table 3.3). One variant is novel and the other is very rare (MAF of 0.00006). A possible explanation is that maybe the disease in this family is not caused by a 'one hit' variant but rather a combination of variants. Perhaps if variation does not occur in all closely linked genes, plausibly a less severe phenotype may be observed, an occurrence that warrants further investigation.

Table 3.3 Novel variants in genes located closely on the same chromosome. Variants were identified in all affected family members and absent in the unaffected one.

Chr	Chr position	Gene	Variation type	Translation impact	Protein variant	dbSNP rs Id	MAF
1	32164127	COL16A1	SNV	Missense	p.T116M	rs34091659	0.0056
1	32204991	BAI2	SNV	Missense	p.P805T	**Novel**	-
1	35575933	ZMYM1	SNV	Synonymous	p.P282P	rs374134267	-
1	35863125	ZMYM4	Insertion	in-frame	p.K1060	**Novel**	-
6	151917596	CCDC170	SNV	Missense	p.H532Y	rs201625561	0.0007
6	152469200	SYNE1	SNV	Missense	R8319Q	rs148008634	0.0006
6	152749340	SYNE1	SNV	Splice loss	p.R1666K	rs111428582	-

3.3.2 IVA identifies novel and rare variants in affected family members

Heterozygous non-synonymous SNVs, insertion/deletions and splice-site variants that were detected in the affected family members and absent in the unaffected member as well as predicted to damaging using SIFT (Ng and Henikoff, 2003), Poly-phen (Adzhubei et al., 2013), Mutation tester (Schwarz et al., 2010) and Snpeff (Cingolani et al., 2012) were prioritised. This reduced the number of plausible candidate pathological variants to 13 in 12 genes (Table 3.4). Of these 3 were novel variants (BAI2 [p.P805T], ZMYM [4p.K1060], KCNN3 [p.Q77_Q80dup]). Of the novel variants, 2 were insertion mutations and one was a SNV. Interestingly, of the 2 splice-site mutations ([p.R1666K, [p.Y96C]) identified, one of them had no MAF reported even though it had a reported snp-id number. The remaining mutations were rare SNVs with MAF < 1%.

Table 3.4 Novel variants identified using IVA. Variants were identified following heuristic filtering as outlined in section 3.2.2.

Chr	Chr position	Gene	Variation type	Translation impact	Protein variant	dbSNP rs Id	MAF
1	32164127	COL16A1	SNV	Missense	p.T116M	rs34091659	0.0056
1	32204991	BAI2	SNV	Missense	p.P805T	**Novel**	-
1	35863125	ZMYM4	Insertion	in-frame	p.K1060	**Novel**	-
1	43919081	HYI	SNV	Splice Loss	p.Y96C	rs14236920	0.00082
1	154842199	KCNN3	Insertion	in-frame	p.Q77_Q80dup	**Novel**	-
6	152469200	SYNE1	SNV	Missense	R8319Q	rs148008634	0.0006
6	152749340	SYNE1	SNV	Splice loss	p.R1666K	rs111428582	-
13	110845216	COL4A1	SNV	Missense	p.R476W	rs369960952	0.0002
16	75269325	BCAR1	SNV	Missense	p.R281H	rs16957558	0.296
16	77246091	SYCE1L	Insertion	Frameshift	p.E164fs	rs371551639	0.0018
19	8145928	FBN3	SNV	Missense	p.R2471H	rs3848570	0.0222
19	10395208	ICAM1	SNV	Missense	p.P352L	rs1801714	0.00072

3.3.3 Statistical variant prioritisation identifies novel variants identical to IVA

To evaluate the probability that a variant is statistically significant in affected compared to unaffected individuals, an analysis using all variants identified in this family was undertaken using VAAST. Interestingly, all variants that were prioritised in IVA were identified in the top 30 genes that were also significant. In fact, BAI1 with a score of 69.14 was ranked first overall, followed by COL4A1 which was assigned a score of 62.50 and BCAR1 with a score of 52.71 were both ranked in the top 5 of VAAST's prioritised hits (Table 3.5). A significant overlap in the genes prioritised using IVA and the top 30 hits from VAAST provided more evidence that a potential causative gene (s) was likely to be amongst these genes.

Table 3.5 Statistical variant prioritisation. Variants that are shared by affected family members and absent in the unaffected family member were sought.

Chr Number	Chr position	Gene	Score	Adjusted p-value	Protein variant
1	32164127	BAI2	69.14	3.73e-16	p.P805T
13	32204991	COL4A1	62.50	3.70e-16	p.R476W
16	75269325	BCAR1	52.71	3.70e-16	p.R281H
19	8145928	FBN3	38.14	0.0056	p.R2471H
1	32164127	COL16A1	31.52	0.001	p.T116M
6	151670287	AKAP12	26.207	0.0052	p.P254L
19	10395208	ICAM1	25.34	0.0004	p.P352L
1	43919081	HYI	25.16	0.000686	p.Y96C
6	152469200	SYNE1	25.06	0.00000762	R8319Q

3.3.4 Genes predicted to have a direct link to End-stage renal disease identified

VarElect was utilised to establish if there are any links between the prioritised genes and the disease of interest. End-stage renal disease, kidney and renal disease were the three terms used to infer the links.

Table 3.6 Genes predicted to have a direct link to End-stage renal disease. The score shows the gene that has more evidence connecting it to renal disease.

Gene	Score
ICAM1	9.07
COL4A1	8.44
COL16A1	8.80
SYNE1	7.77

Four genes were predicted to have direct links to ESRD and all of them had scores above 5.00 (Table 3.6). These were prioritised as genes of interesting genes.

3.4 Potential disease causing genes identified in all affected family members

Table 3.7 Prioritised potential disease causing genes. Variants in these genes segregated with disease in all affected family members and were absent from one unaffected family member. The variants were prioritised using a combination of heuristic filtering and statistical probabilistic variant analysis.

Chr	Chr position	Gene	Protein variant	dbSNP rs Id	Translation impact	MAF	Polyphen	Gerp	SIFT Prediction
13	110845216	COL4A1	p.R476W	rs369960952	Missense	0.0002	0.996	4.87	Damaging
1	32164127	COL16A1	p.T116M	rs34091659	Missense	0.0056	1.000	3.06	Damaging
19	10395208	ICAM1	p.P352L	rs1801714	Missense	0.0007	0.997	3.78	Damaging

*PolyPhen2 score >0.85 the variant is predicted to be damaging.

* GERP score is a measure of evolutionary conservation and a score > 2.5 is conserved.

* MAF <0.1% was considered very rare

Table 3.7 shows a list of high impact variants identified using a combinatorial approach to variant prioritisation. To infer potential functional significance of prioritised variants, I applied a protein variation effect analyzer PROVEAN v1.1.3 (*http://provean.jcvi.org*). PROVEAN human genome variants tool provides predictions for a given list of human genome variants as well as accessory information (dbSNP rs IDs, gene description, PFAM domain, GO terms, etc.) and is able to make predictions for any type of protein sequence alteration, including single or multiple amino acid substitutions, deletions, and insertions. Also, the 3 missense variants were all predicted to be damaging using both SIFT and PolyPhen (Table 3.7). The variants also occurred in conserved genomic locations across a number of vertebrate species inferred with a GERP score greater than 2.5 (Table 3.7). All variants were very rare with a MAF< 0.1 % (Table 3.7). The same variants had a VAAST score greater than 25 and a p-value less than 5% (Table 3.5). Also, these variants had an endogenous score greater than 5 linking them to ESRD (Table 3.6). To illustrate the potential involvement of these genes in renal diseases (Table 3.7), rigorous functional analyses were undertaken. Subsequent sections provide more detail of the plausible mechanisms and functions these genes are involved in and how they relate to the phenotype under investigation.

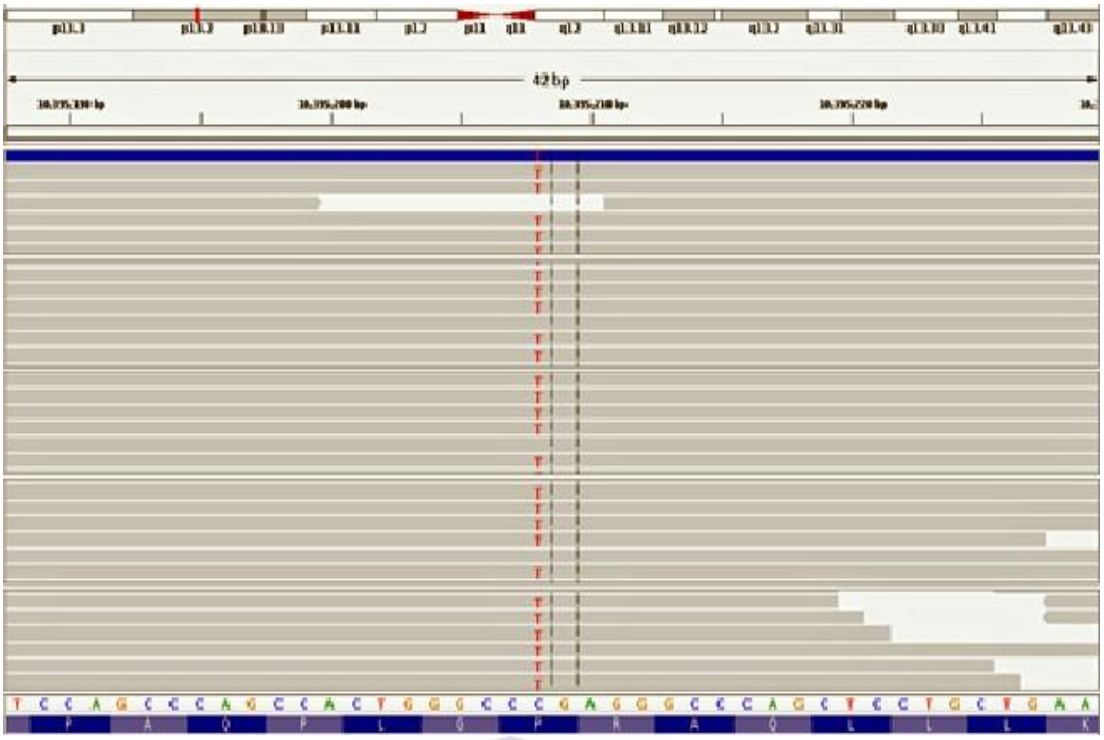


Figure 3.4 ICAM1 [p.P352L] variant visualisation. The C>T allele change in ICAM1 is seen all the 5 affected family members.

COL4A1 one of the genes identified as disease causing (Table 3.7) has three major domains: an amino-terminal 7S domain, a central triple-helix-forming (collagenous) domain and a carboxyterminal non-collagenous (NC1) domain (Figure 3.5). The 7S domain participates in inter-molecular cross-linking and macromolecular organization. The collagenous domain constitutes the majority of the protein and consists of long stretches of glycine repeats. The NC1 domains are globular domains responsible for the initiation of heterotrimers assembly (Figure 3.7).

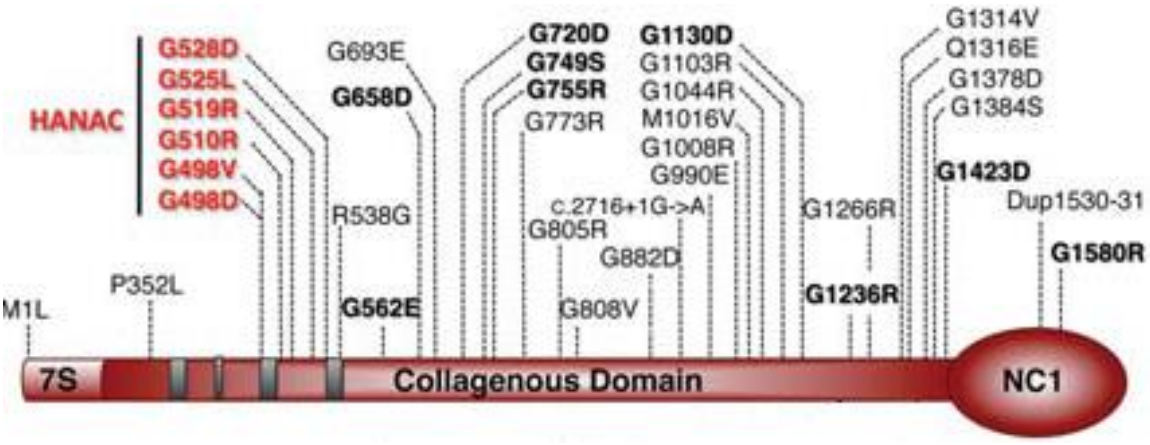


Figure 3.5 Human schematics of the distribution of COL4A1 mutations. Hereditary angiopathy with ***nephropathy***, aneurysms and muscle cramps mutations are shown in red.

3.4.1 Candidate genes are involved in increased glomerulus injury, renal damage and renal failure

Ingenuity pathway analysis was carried out to infer if the prioritised genes are involved in any pathways related to kidney disease.

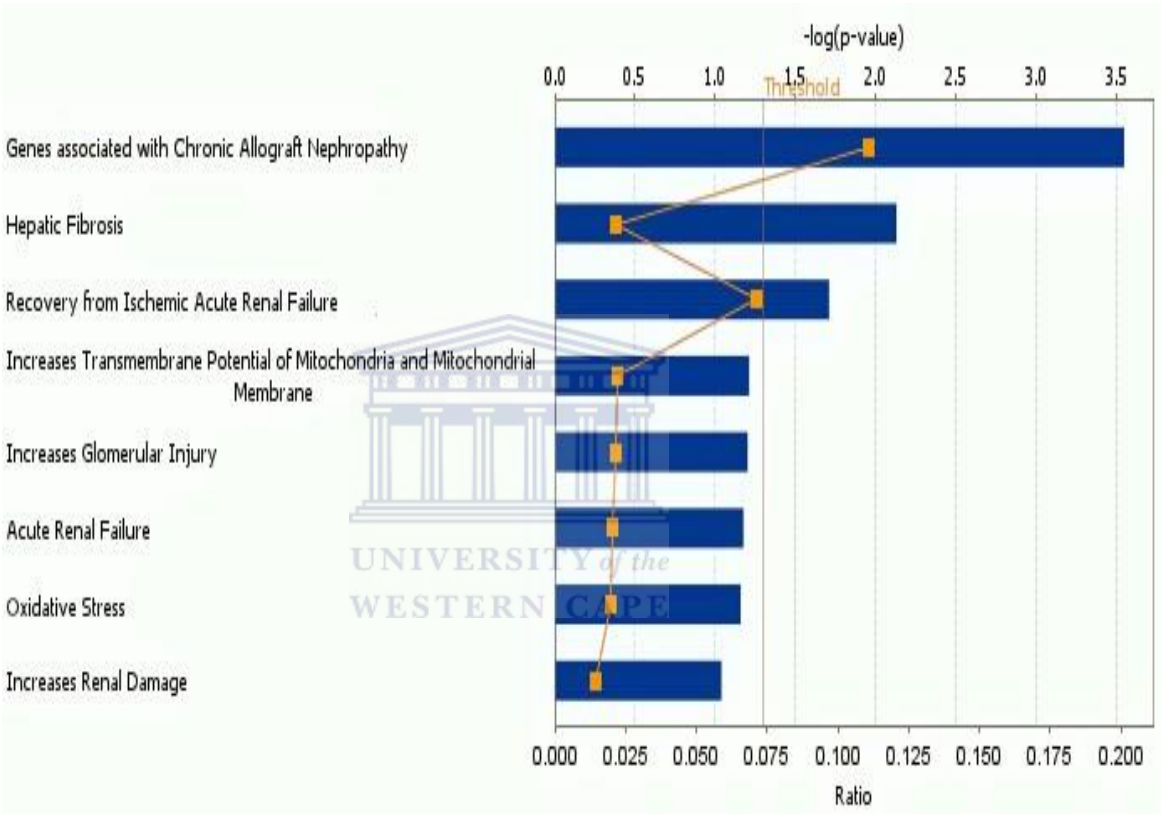


Figure 3.6 Pathways enriched from the prioritised candidate genes. The pathways were generated using a list of variants prioritised using Ingenuity pathway analysis.

Figure 3.6 shows increased glomerular injury, increased renal damage and acute renal failure as some of the enriched toxic pathways. Three genes, COL4A1, COL16A1 and ICAM1 are involved in these kidney disease related pathways. The same 3 genes were identified as being directly linked to End-stage renal disease in Table 3.6. Also, COL4A1 was amongst the top 5 hits from VAAST analysis (Table 3.5). The toxicity pathways identified were all statistically significant; chronic allograft nephropathy (p-value 0.0000281), Hepatic fibrosis (p-value 0.000739), increased glomerular injury (p-value 0.00626) and Acute renal failure (p-value 0.00192) (Table 3.8). Further

functional analysis of potential candidate genes identified embryonic development, tissue development, cell to cell signalling interaction and cell morphology as some of the enriched functions (Table 3.8).

Table 3.8 Molecular, cellular and System development functions enriched.

Function	P-value
Physiological System Development and Function	
Embryonic Development	0.00014
Hematological System Development and Function	0.00014
Organismal Functions	0.00399
Tissue Development	0.00070
Molecular and Cellular Functions	
Cell Death and Survival	0.00392
Cell Morphology	0.00040
Cell-To-Cell Signaling and Interaction	0.00040
Cellular Assembly and Organization	0.0040
Cellular Compromise	0.0042

3.4.2 Candidate genes are involved in interstitial fibrosis

A kidney biopsy that was done on one of the affected family members showed interstitial fibrosis. Based on the findings of this kidney biopsy, I investigated if any of the potential causative genes are involved at any stage of interstitial fibrosis.

.

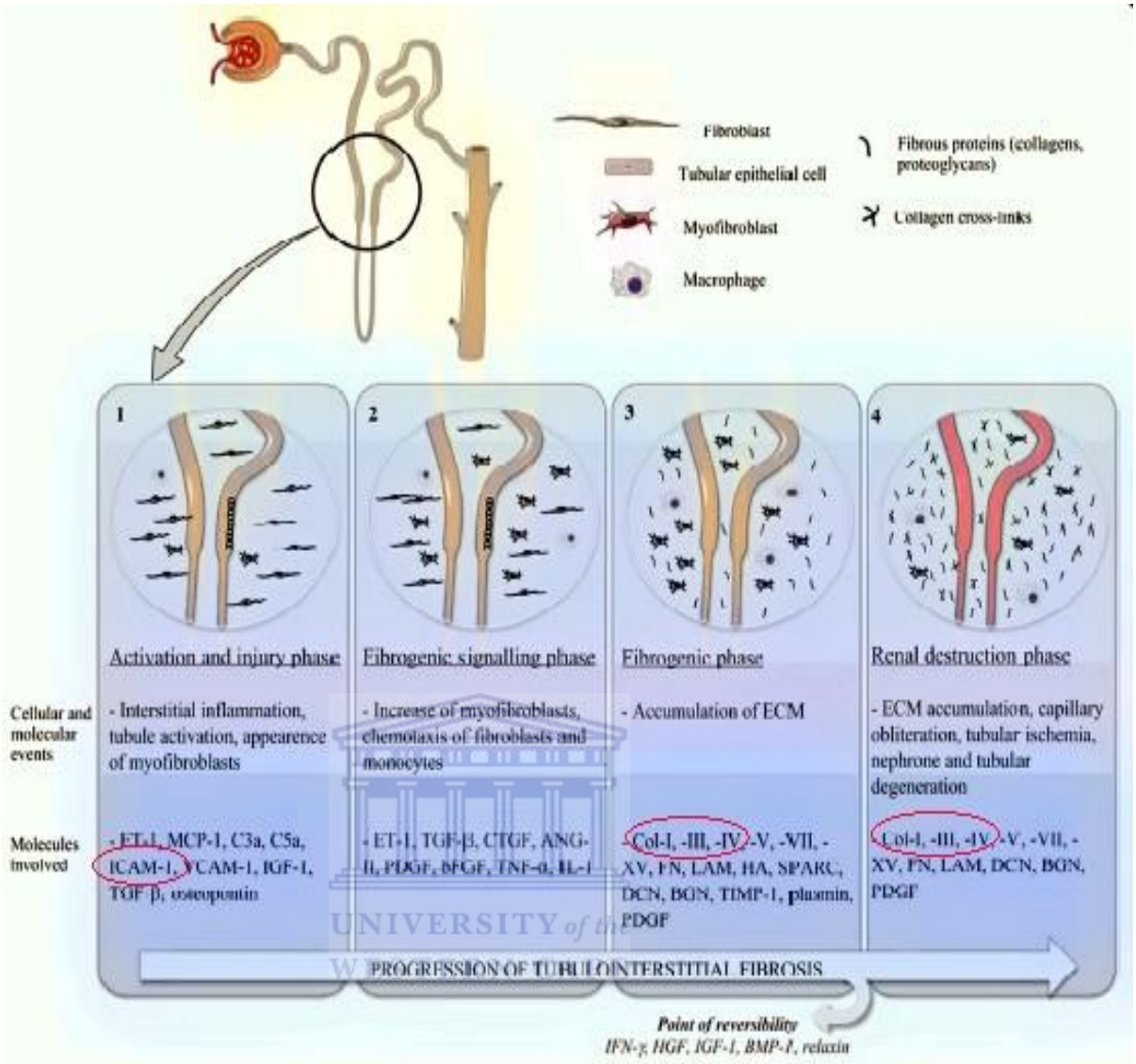


Figure 3.7 Progression of renal interstitial fibrosis towards End stage renal disease. Fibrogenesis starts with an initial tissue injury that causes inflammation as the physiological host defense response. When this response becomes uncontrolled and sustains itself with continuous production of chemotactic cytokines, inflammation does not resolve and can create the optimal microenvironment for tissue fibrogenesis. ICAM1 and COL4A1 are involved in stage 1, 3 and 4 of interstitial fibrosis progression (Genovese et al., 2014b).

ICAM1 and COL4A1 are involved in different stages of renal interstitial fibrosis progressing towards end stage renal disease (Figure 3.7). ICAM1 is involved in the activation phase of fibrogenesis while COL4A1 is involved in the accumulation phase and the more severe renal destruction phase. Figure 3.8 shows the plausible mechanism through which COL4A1 interact with extra cellular matrix components (ECM) that are involved in interstitial fibrosis.

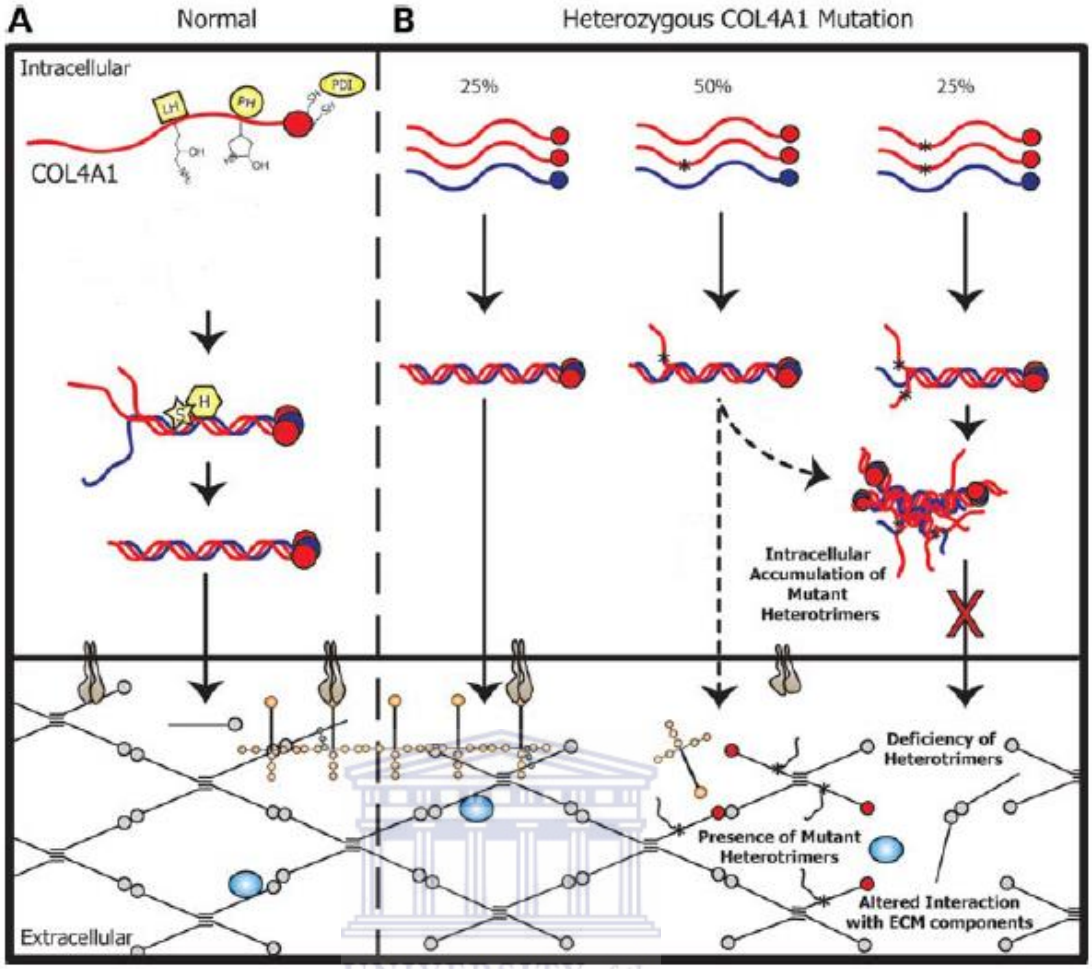


Figure 3.8 Schematic representation of COL4A1 biosynthesis and interaction with extra cellular matrix components. (A) Collagen proteins undergo extensive post-translational modifications and assemble into heterotrimers for secretion into the ECM where they polymerize into a network and interact with other extracellular and membrane bound molecules such as lysyl hydroxylase, prolyl hydroxylase and protein disulphide isomerase (B) shows COL4A1 mutations of which 25% of heterotrimers formed will be normal, 50% will incorporate one mutant COL4A1 protein and the remaining 25% will incorporate two mutant COL4A1 proteins. The mutations could directly or indirectly alter interactions with signaling molecules such as BMPs (represented as blue circles) or cell-surface receptors such as integrins (represented as grey structures), which can in turn lead to intracellular signaling defects (Kuo et al., 2012).

3.4.3 Candidate variants are conserved across species

Evolutionary conservation analysis was undertaken to establish if candidate variants occur in conserved genomic locations across a number of species. If a variant occurs in the conserved region it is likely to have a functional impact that might be deleterious.

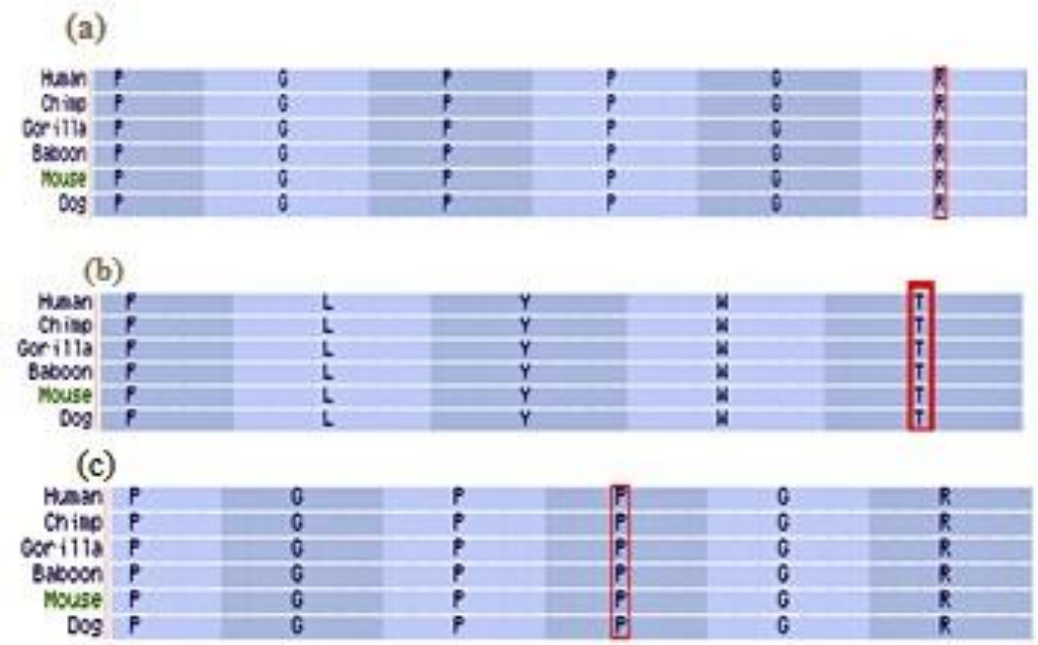


Figure 3.9 Evolutionary conservation of mutation identified in affected family members. Multi-species sequence alignment shows that COL4A1, ICAM1 and COL16a1 mutations are highly conserved across species (a) Shows the evolutionary conservation a p.Arg476Trp missense variant in COL4A1 gene (b) Evolutionary conservation of a p.Thr116Met missense variant in COL16A1 gene (c) Shows the evolutionary conservation of a p.Pro352Leu missense ICAM1 gene.

A p.Pro352Leu amino acid substitution was identified at position 476 in the COL4A1 gene. This is an N-acetylmuramoyl-L-alanine amidase glycosylation site. This position is is a highly conserved among different species from a dog to human, suggesting its structural and functional importance across many species. It also has a low probability of substitution with BLOSUM score 3. In COLl16A1 a p.Thr116Met amino acid substitution was identified at position 116. This site was also found to be highly conserved across many species. Similarly, a p.Arg476Trp amino acid substitution was observed in the conserved site of the ICAM1 gene.

3.4.4 Protein-protein interaction networks and gene co-expression analysis of candidate genes

A gene co-expression network was constructed for each candidate gene by looking for genes which show a similar expression pattern across all affected family members. Identifying genes that are co-expressed is of biological interest as co-expressed genes may be controlled by the same transcriptional regulatory mechanism, are functionally

related, or the genes are members of the same pathway.

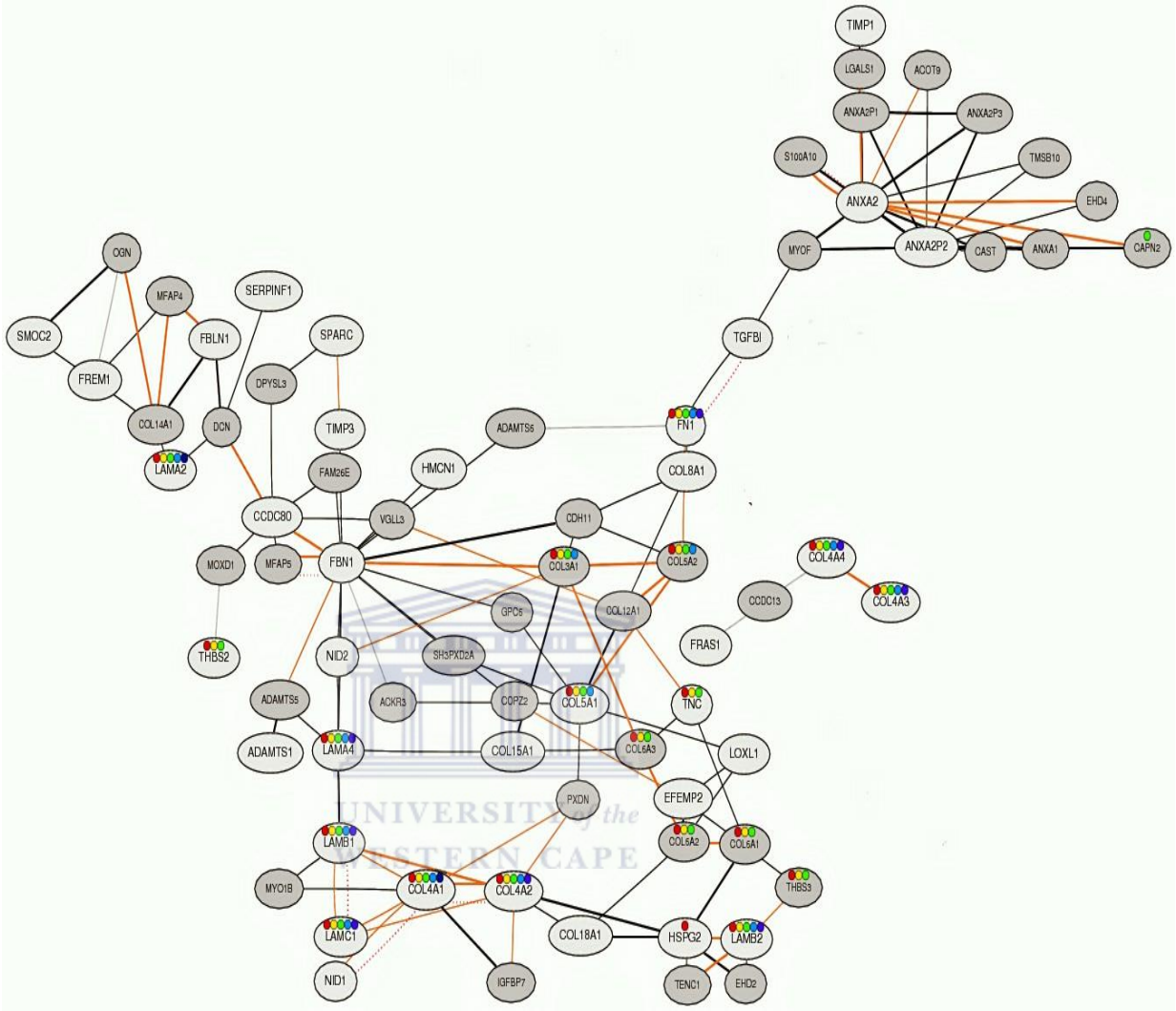


Figure 3.10 COL4A1 co-expression analysis. The analysis was performed for co-expression in the basement membrane.

Figure 3.10 shows that in the basement membrane COL4A1 co-expresses predominantly with other collagen genes (col3a1, col4a2, col15a1, col15a2). This pattern is also consistent with the protein-protein interaction network (Figure 3.11) which shows that most collagen proteins interact, highlighting the fact that collagen genes together are important for the basement membrane function and organisation. Also, NID1and ITGB1 are observed in both the gene co-expression analysis (Figure 3.10) and protein- protein interaction analysis (Figure 3.11).

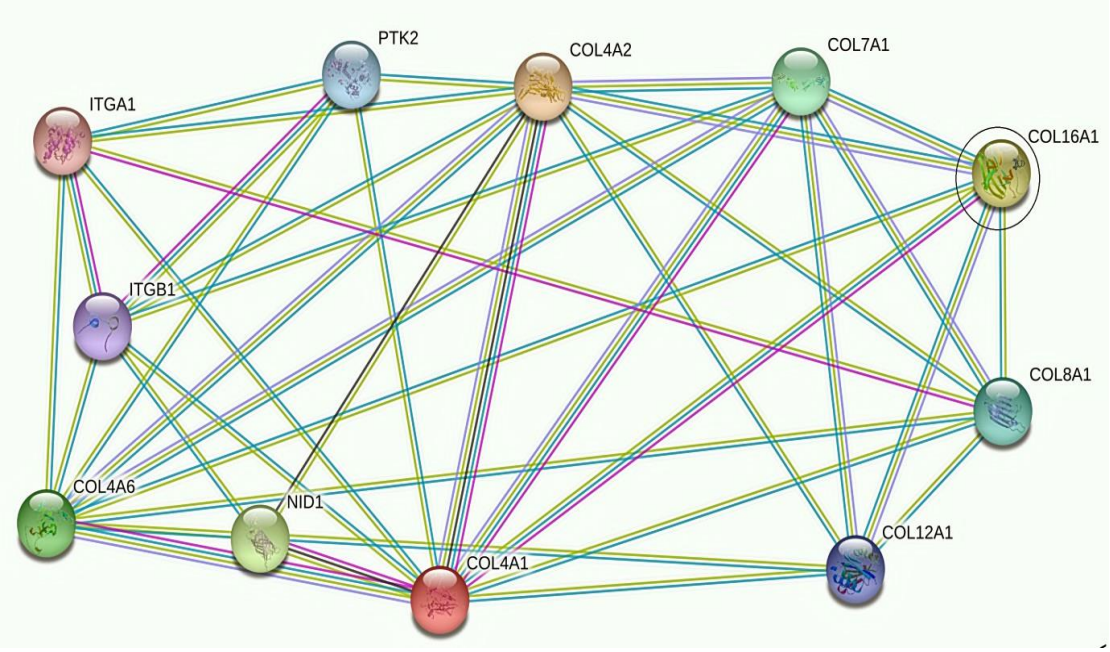


Figure 3.11 Protein-protein interaction networks. . This analysis was done using STRING. COL4A1 was used as an input and a list of genes whose proteins interact with this gene were shown. Different colours indicate the strength of the interaction

Collagen, type IV, alpha 1 is the major structural component of the glomerular basement membranes (GBM) which form a key filtration mechanism of the kidneys. COL4A1 proteins interact significantly with other collagen proteins for example COL4A6 and COL8A1, illustrating the importance of these proteins in the structural make-up of the glomerulus. Interestingly, COL4A1 proteins interact directly with COL16A1 proteins (Figure 3.11). COL16A1 is one of the genes that were implicated as potentially disease causing (Table 3.7). The same two genes were involved in glomerulus injury (Figure 3.6). Also, other genes such as ITGA1 and PTK2 form protein products that interact directly with COL4A1 and COL16A1. Thus, the collagens genes are interesting to follow up as they form are integral structural components of the glomerulus.

3.5 Protein structure modelling

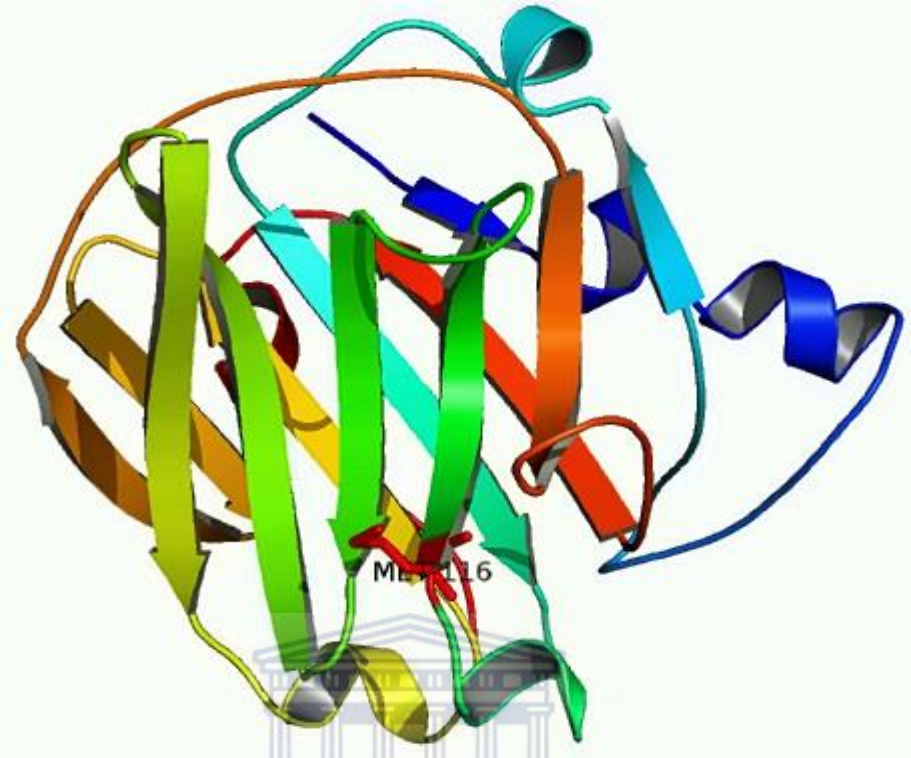


Figure 3.12. Col16A1 3D model with a p.T116M variant introduced. The green colour shows the beta sheet while the blue and orange are the alpha helix structures. The red shows the position of the variant on the beta sheet.

The 3D structure of the plausible causative genes was modelled using Swiss modeller (Schwede et al., 2003). Also, Pymol (*www.pymol.org*) was used to introduce the mutation and visualise it (Figure 3.12). The p.T116M mutation in Col16a1 occurs in the ankyrin repeat domain which results in the change from Threonine (T) to Methionine (Met) amino acid residue at position 116. This is a change from medium size hydrophilic amino acid (Ther) to medium size and hydrophobic (Met) (Figure 3.13). The beta sheet and the alpha form a very complex structure and any mutation that occurs in this region of the protein may result in adverse consequences such as loss of interaction with other residues in the region, can cause other residues in that region to be invisible and may alter the overall structure of the protein. Given that the mutation occurs in the conserved region (Figure 3.9), further studies which are beyond the scope of this thesis needs to carried out to investigate if the mutation occurs in a binding pocket and the extent to which the mutation alters the overall structure of the protein. Also, the two amino acids have different molecular weights

(MW) with Ther having a MW of 101.11 and met a MW of 131.19. The 3D structure of ICAM1 with a p.P352L amino acid change is shown in appendix F. The Col4a1 structure could not be modelled as there was no suitable template to perform the modelling.

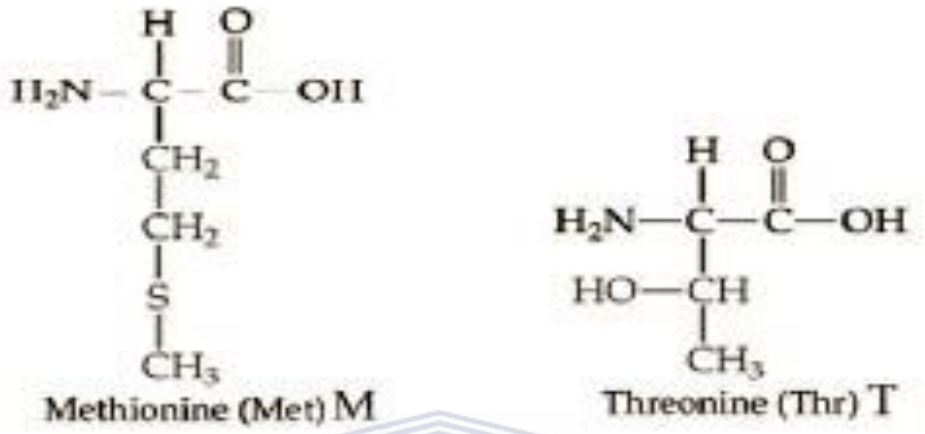


Figure 3.13.Molecular structure of the amino acid residues. The molecular weight of M is 131.19 while that of T 101.11.

3.6 Discussion

Options currently available for clinical genetic testing include next generation sequencing of a selected panel of candidate genes, WES, and whole genome sequencing (Liew et al., 2013; Stitziel et al., 2013; Worthey et al., 2010). In clinical settings, targeted panels are being leveraged because of their relatively modest cost and the detail with which the targeted regions are analyzed which allow reliable copy number and microsatellite estimation. A major drawback of this approach is that if the panel used does not include the gene responsible in an affected person, the variant will not be detected (Taylor et al., 2015). This is a significant problem in the current era where new genes are being implicated in mendelian kidney diseases each year.

On the other hand, WES yields data on almost all known genes, including those not initially considered candidates for the phenotype being investigated. This allows the possibility for a molecular or genetic diagnosis to be made where detailed phenotype data such as kidney biopsy are lacking, and variants that predict disease may be

implicated. In this study, 5 affected family members posed a diagnostic challenge in kidney disease, presenting with elevated serum creatinine with rare and surprising absence of proteinuria and hematuria. Also, all affected family members rapidly progress to early onset ESRD, requiring kidney dialysis or transplantation. Thus, the genetics underlying ESRD in this family were investigated. Sequencing 5 affected and one unaffected family member using WES, followed by a stepwise variant filtering strategy and probability variant prioritization, we identified 3 strong candidate disease-causing variants: p.Arg476Trp in collagen IV (COL4a1), p.Thr116Met in collagen 16 (COL16a1) and p.Pro352Leu ICAMA1 gene (Table 3.6).

Two of the high impact variants prioritized were in the collagen genes (Table 3.7). Collagen type IV is composed of alpha (α) chains that fold in a triple helix and, by binding with other collagen type IV molecules form the meshwork conformation typical of the basement membrane. The α chains are the most expressed in the adult glomerular basement membrane GBM (Genovese et al., 2014b). Large indels, rearrangements, splicing and nonsense mutations in collagen IV gene have been previously associated with severe consequences of ESRD occurring in some cases before the age of 20, whereas missense mutations involving the collagenous domain are responsible for renal diseases such as Alport Syndrome (Gross et al., 2002; Yao et al., 2012). Dysfunction of the collagen IV genes disrupts proper heterotrimer formation causing a failure to deposit normal collagen matrix in the glomerular basement membrane resulting in the disruption of the kidney filtration barrier (Figure 3.7). In these cases, glomerular dysfunction is the common cause of end stage renal disease (Gipson et al., 2006). In the list of prioritized genes, pathway analysis performed in IPA identified COL4A1 and COL16A1 as genes involved in increased glomerular injury, increased renal damage and acute renal failure (Figure 3.6). This shows the importance of the collagen genes in forming the main structural components of the glomerulus, which is a key filtration component of the kidney.

Based on a kidney biopsy for one of the affected family members, interstitial fibrosis was observed. Interstitial fibrosis is the common endpoint of end-stage kidney disease leading to kidney failure (Genovese et al., 2014). Interstitial fibrosis is the strongest indicator of disease progression, even when the primary disease is of unknown etiology (Genovese et al., 2014). The development of novel, non-invasive, fibrosis-

specific biomarkers, reflecting morphological tissue changes at early stages and predicting the evolution of renal fibrosis, would be of vital importance to delay progression of renal disease to ESRD, which is a more severe form of kidney disease (Genovese et al., 2014). COL4A1 and ICAM1, two genes identified as potentially disease-causing, are involved in different stages of interstitial fibrosis (Figure 3.7). Collagens constitute the main structural element of the interstitial extra cellular molecules, providing tensile strength, regulating cell adhesion, support, cell migration and tissue development (Rozario and DeSimone, 2010). Elevated Collagen IV levels in patients with various nephropathies have been found to correlate with the extent of interstitial fibrosis in kidney biopsies (Soylemezoglu et al., 1997). Also, collagen IV levels evaluated in kidney transplant patients correlated with the extent of interstitial fibrosis (Teppo et al., 2003).

Furthermore, in another study of patients with different chronic kidney disease stages subjected to kidney biopsy, collagen IV levels correlated with elevated serum creatinine and eGFR. Similarly, all sequenced affected family members in this study presented with elevated serum creatinine (Ghoul et al., 2010). Providing more functional evidence to corroborate the implication of COL4A1 as a potentially disease causing gene. Thus, by profiling these collagen genes, one may plausibly understand mechanisms underlying rapid progression of chronic renal disease to ESRD in this family.

Studies using targeted panels of COL4A3/A4/A5 for the genetic diagnosis in patients suspected of having a Type IV collagen-related nephropathy identified a likely pathogenic mutation in 55% and 83.2% patients tested (Fallerini et al., 2014; Morinière et al., 2014). Mutations in genes encoding α chain of type IV collagen could lead to dysfunction of glomerular basement membrane (BM) leading to the development of human disease in the eye, kidney and ear (Deltas et al., 2013) . Once the α chain is missing, the formation of the normal collagen IV is disrupted in BM of glomerulus, ear, eye, and lung, which could lead to structural and functional defects (Frasca et al., 2004). This is supported by the immunohistochemical finding of frequent loss of $\alpha 3$, $\alpha 4$, and $\alpha 5$ signals in the GBM of Alport syndrome patients (Gross et al., 2002; Yao et al., 2012). Unfortunately, no formal testing was undertaken

in this family to establish clinical features of either sensorineural hearing loss or related ocular abnormalities (although the proband, now deceased, did present with intermittent periods of impaired vision). This would have added vital information since the mutations identified in collagen genes may have potential involvement in the eyes and ears. Collagen IV genes have been implicated in several studies seeking to establish the genetics underlying rare renal phenoptyes (Chatterjee et al., 2013; Pierides et al., 2009).

Another gene that was prioritized in this study is ICAM1. Intercellular adhesion molecule ICAM plays a crucial role in the pathogenesis of primary kidney disease and progression to end-stage renal disease (ESRD) (Khazen et al., 2007; Ong and Fine, 1994; Vleming et al., 1999). Intercellular adhesion molecule-1 (ICAM1) is a leukocyte adhesion molecule, which is expressed at high levels in the kidney on the endothelial cells and interacts with integrins (Mclaren et al., 1999). It is involved in leukocyte adhesion, recruitment and also enhances the activation of T helper cells (Mclaren et al., 1999). The recruited leukocytes in turn release cytokines, such as platelet-derived growth factors. These transform growth factors and fibroblast growth factors, which stimulate extracellular matrix production by interstitial cells such as fibroblasts causing interstitial fibrosis (Figure 3.7). Adhesion molecules provide signals for activation and recruitment of effector cells, leading to graft infiltration by host T-cells, which are important to allograft rejection (Khazen et al., 2007). Several polymorphisms in ICAM1 have been discovered to be associated with diseases such as acute renal allograft rejection (Khazen et al., 2007; Mclaren et al., 1999). Interestingly, some of the family members sequenced have experienced kidney transplant failure and results obtained from pathway analysis showed increased glomerular injury and chronic allograft failure as some of the enriched toxic pathway (Figure 3.4). In a study conducted on 258 ESRD patients and 569 ethnically matched controls (Ranganath et al., 2009). ICAM1 polymorphisms investigated were found to be significantly different in ESRD patients when compared with controls (P ¼ 0.0001; OR ¼ 5.5, 95% CI ¼ 3.9–7.7 and P < 0.0001; OR ¼ 3.8, 95% CI ¼ 3.1–4.7) (Ranganath et al., 2009). These results demonstrated that various SNPs in the ICAM1 gene may be considered as genetic variants that influence susceptibility to ESRD (Ranganath et al., 2009).

This study showed that exome sequencing is a fast, sensitive, and relatively low-cost method of identifying gene(s) responsible for rare familial end stage renal disease. The identified COL4A1 and COL16A1 genes broaden the genotypic spectrum of collagen mutations associated with renal diseases and have implications for genetic diagnosis, therapy, and genetic counseling in this family. Also, this study emphasizes the role of molecular diagnosis in aiding the phenotypic characterization of different kidney diseases and selection of appropriate treatment modalities. Finally, the results show that WES is a powerful diagnostic tool that can complement invasive procedures such as renal biopsy and provide a diagnosis in patients with familial kidney disease, particularly when clinical information is limited or non-specific.

3.7 Conclusion

Next generation sequencing techniques and bioinformatics approaches applied in this study identified 3 very rare pathogenic missense variants in COL4A1, COL16A1 and ICAM1 segregating with an unexplained inherited kidney disease in 5 affected family members. These findings highlight the clinical range of collagen related nephropathies and may resolve diagnostic difficulties arising from lack of and uninformative clinical and histological findings, allowing appropriate treatment advice to be given. To our knowledge this is the first application of this approach to unravel the genetics underlying familial ESRD in a South African population and highlights the need to study further the genetics of ESRD in African populations.

4 Clinical databasing

Abstract

Lack of systematically collected clinical data on disease characteristics and long-term outcomes in patients with CKD and its risk factors is one of the major problems which hamper the fight against CKD and subsequently ESRD. To help address this problem, in collaboration with clinicians at Groote Schuur hospital, Cape Town, South Africa, we have set up a multicentre clinical registry. The clinical registry will collate demographic, epidemiological and basic clinical data of patients. This information can broaden our knowledge on patient diagnosis, treatment, clinical patterns and outcomes. Overall, the clinical registry establishes a platform to provide resources for future clinical and genomic studies in Africa.

4.1 Introduction

In collaboration with nephrologists at University of Cape Town Medical School (Groote Schuur Hospital), we identified a lack of systematically collected clinical data on disease characteristics and long-term outcomes in patients with CKD and its risk factors as one of the major problems in the fight against CKD and subsequently ESRD. The lack of such systematically collected data presents a gap that needs to be urgently bridged as a crucial initial step towards confronting the burden of CKD and its risk factors, especially in SSA (Singh et al., 2012). Reliable data that can be drawn from these clinical databases might assist policy makers in low income countries to formulate strategies and interventions that can be used to improve diagnosis, treatment and management of CKD and its risk factors, which may eventually lead to improved patient outcome (Okpechi et al., 2010).

A clinical database is any systematic compilation of data for the purpose of health care planning, implementation and evaluation in a well-defined population. Clinical databases may contain a large variety of data from different domains, such as patient visits, test results, laboratory reports, diagnoses, therapy, medication, and procedure (Sam Lim et al., 2009). Also, they may have different purposes which may include patient management, electronic patient records, clinical research, and quality control (Saghir et al., 2007). Clinical databases are also a valuable complement to randomized

controlled trials in determining real-world outcomes in the practice of medicine (Brooke and others, 1974). They do not generally have a lot of restrictive inclusion-exclusion criteria; neither do they specify what therapy the health care provider must adhere to. Clinical databases can be used to evaluate outcomes ranging from the history of a disease, to disease presentation, prognosis, to the safety of drugs and effectiveness of therapies (Singh et al., 2012). Also, epidemiological research on disease occurrence and distribution, disease risk or etiology and disease prevention can be done using data from clinical databases (Sam Lim et al., 2009).

In this Chapter, I report a clinical database that I designed for Systemic Lupus Erythematosus (SLE), one of the major risk factors for ESRD in SSA. The choice of SLE was motivated by the fact that the prevalence of this disease is high amongst CKD patients that are being treated by Nephrologists at Groote Schuur Hospital in Cape Town, South Africa. SLE is a disease which requires histopathological diagnosis in order to treat properly and manage (Arogundade et al., 2011); however, reliable statistics that are required to elucidate epidemiological patterns of SLE in SSA are difficult to obtain and are largely undetermined (Tiffin et al., 2013). On the other hand, the observed low incidence rate of SLE in Africa may be attributed to under diagnosis, low disease recognition within primary health care facilities but more importantly the often neglected limited access to diagnostic tools of which clinical databases are an essential part. In contrast, incidences of SLE have been studied comprehensively in European, Asian, African-American, Hispanic and Caribbean populations (Danchenko et al., 2006). Therefore, it has become increasingly imperative that a formal structured way of storing clinical data for patience with SLE be sought in order to better understand the presentation, diagnosis, prognosis, therapies and treatment outcome of SLE patients (Tiffin et al., 2013). Hence, setting up of a clinical database can go a long way in bridging the gap and provide the much needed data (Lu et al., 2010). Going forward, such a valuable clinical research resource will also become important in African genomic studies for both hypothesis-led and hypothesis-generating research approaches that maybe undertaken to better understand the causes, prognosis, management and outcomes of SLE (Villa-Blanco and Calvo-Alén, 2012). Importantly, once data is available in a formally structured and secured database then it becomes easier to analyze this data and provide valuable insights that may also be used to inform allocation of resources, for example health

workers and to also see which treatment regimens are working and for which patients (Villa-Blanco and Calvo-Alén, 2012) . Designing of clinical databases is an area of clinical informatics research which ought to be given some attention and this thesis addresses a part of this problem.

4.2 Methods

Critical factors to consider when designing a clinical database include defining the population to which the findings are meant to apply, formulating a research question, choosing a study design, translating questions of clinical interest into measurable exposures and outcomes, choosing patients for study, determining where data can be found and for how long patients will be followed up. The number of study subjects desired and length of follow-up should be planned in accordance with the overall purpose of the clinical database. The desired study size can be determined by specifying the magnitude of an expected clinically meaningful effect or the desired precision of effect estimates. Study size determinants are also affected by practicality and cost. Once these key design items have been determined, the database design should be reviewed to evaluate potential sources of bias and these should be addressed to the extent that is practical and achievable.

4.2.1 Database construction

The clinical database was designed as a longitudinal multi-center database; we envisaged this being used as a Pan-African clinical registry for SLE patients. The clinical database will be utilized under the African Lupus Genetics Network (ALUGEN), a network of clinicians and researchers in Africa who have an interest in SLE. Given that SSA is a resource limited region, cost effective methods for designing a clinical database were sought. Research electronic data capture (REDCap), a secure, web-based application designed to support standardized collection of research data was utilized in designing the clinical database (Harris et al., 2009). REDCap uses PHP, JavaScript programming and MySQL database engine for data storage and manipulation (Hillyer, 2010; MySQL, 1997; Severance, 2012). REDCap's software and hardware requirements are minimal as it can be run on

Windows, Linux machines and Apache web server environments. The clinical database will be centrally stored and backed up daily and is supported by an experienced team of software developers and statisticians at SANBI.

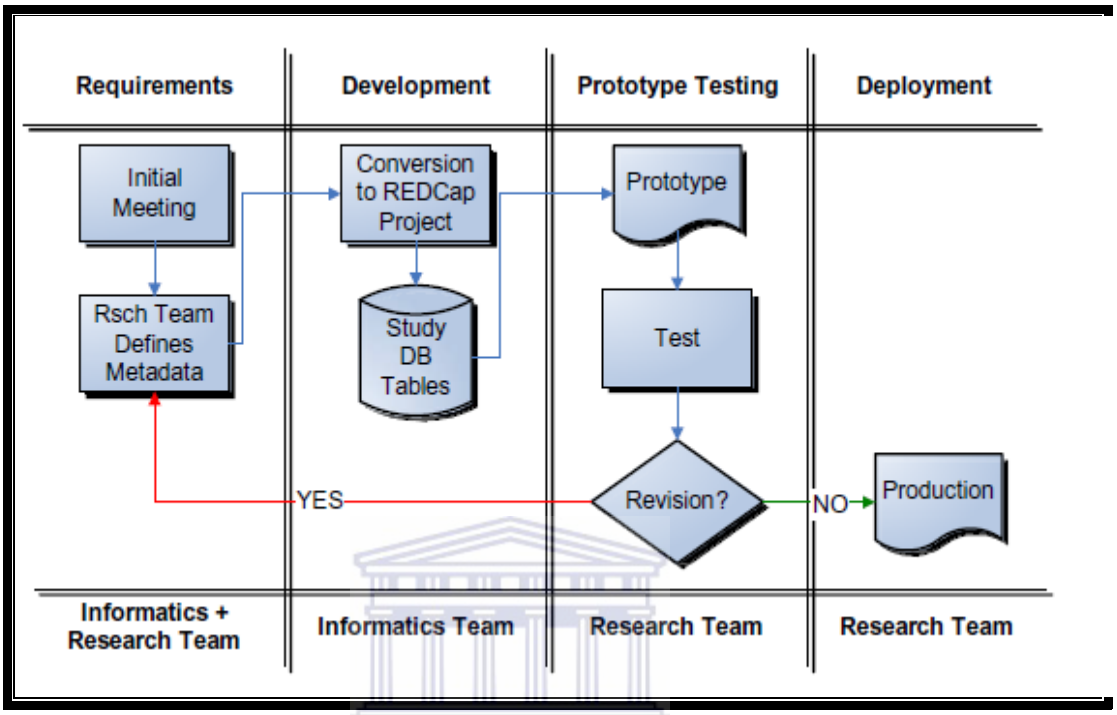
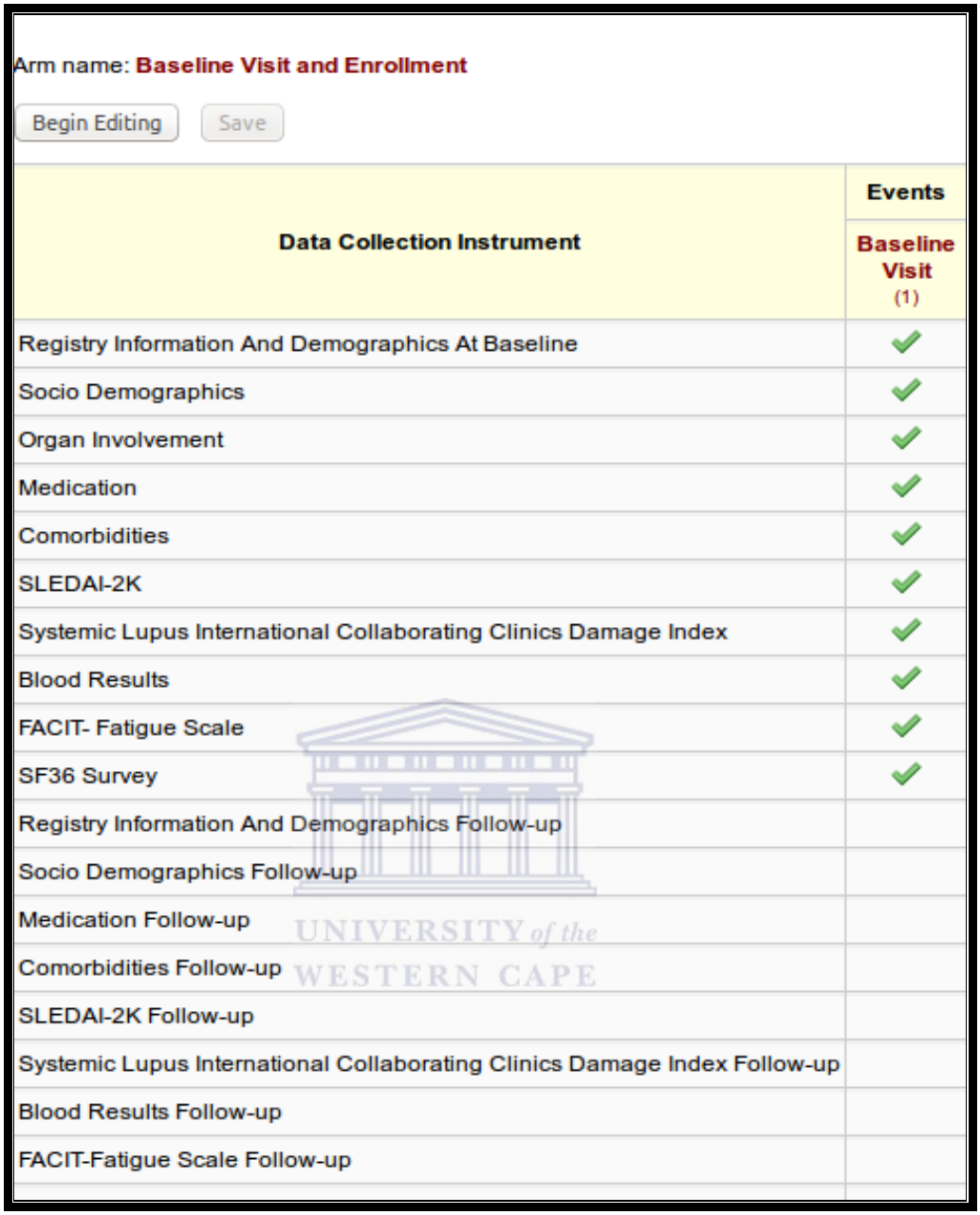


Figure 4.1 A series of steps that are undertaken to design a clinical database using REDCap. The metadata is composed mainly of CRFs that are used for the development phase to create the database tables. Once finished the database prototype is tested to ensure that it adheres to all quality control and security checks. After testing the database can either be redesigned as illustrated by the red arrow or migrated to production mode shown by the green arrow where it can be fully functional.

In a database, an entity is a single person, or place, for example, a patient or a diagnostic test about which data can be stored. In relational database design, each entity is mapped to one or more tables using values of one or more rows to uniquely identify each record. That means that for each entity there exists at least one table. To give users access to the new tables, new forms must be designed and links to these forms must be provided in the user interface. If a table that is already in the database needs to be modified care must be taken not to destroy existing data and not to break any constraints.

Accordingly, user-interface forms must be redesigned to reflect changes e.g., fields that have been added or removed in existing tables. Hence, the database was designed to reflect exactly the information that is on the case report forms (CRFs), as defined

by the research team (Figure 4.1). The CRFs are used to create tables which are stored in a single MySQL database (Figure 4.1). The users of the database will access the database through a web browser and they will only have access to the data entry forms (Figure 4.2). Figure 4.1 shows in detail the steps that are followed in designing a clinical database using REDCap. The database was designed to have two arms: the first arm captures baseline and enrollment patient data. The second arm captures data for follow-up visits. Data collection instruments were assigned according to which arm of the registry they belonged to and this means that the data collection instrument will only be accessible to that arm.



Arm name: **Baseline Visit and Enrollment**

Begin Editing | Save

Data Collection Instrument	Events: Baseline Visit (1)
Registry Information And Demographics At Baseline	✓
Socio Demographics	✓
Organ Involvement	✓
Medication	✓
Comorbidities	✓
SLEDAI-2K	✓
Systemic Lupus International Collaborating Clinics Damage Index	✓
Blood Results	✓
FACIT- Fatigue Scale	✓
SF36 Survey	✓
Registry Information And Demographics Follow-up	
Socio Demographics Follow-up	
Medication Follow-up	
Comorbidities Follow-up	
SLEDAI-2K Follow-up	
Systemic Lupus International Collaborating Clinics Damage Index Follow-up	
Blood Results Follow-up	
FACIT-Fatigue Scale Follow-up	

Figure 4.2 Allocating data entry forms for database arms. Green ticks show the data entry forms to be completed for a baseline visit only.

The user of the clinical database will enter their data through an intuitive secure and accurate user interface (Figure 4.3). Only users with sufficiently assigned privileges will have access to the database and subsequently the data entry forms (Figure 4.2). To ensure high quality and data integrity, each form contains real time field-specific validation. To safeguard against omission of important clinical data during data entry, mandatory fields were set up in the database and upon entry users will be alerted if they did not complete a required field. To improve further the quality of data collection, data fields were populated with drop down and radio boxes (Figure 4.3).

Similarly, additional quality control measures were implemented that would flag the user if a value that is out of range is entered for a customized field. Also, the clinical database has a data export facility that allows users to export their data for external analysis. This allows the database users to export data to a variety of widely used statistical analysis software such as SAS, SPSS, STATA and R. Since it's a multi-center-user clinical database, the data will be stored in such a way that users will only access data for their own group, unless an arrangement is made for extended group access and data sharing, of which prior approval and memorandum of understanding needs to signed. This is meant to ensure that each group returns sole and secure ownership of their clinical data. This system also ensures that ethical constraints that are centre-specific are not violated by users from other centers.

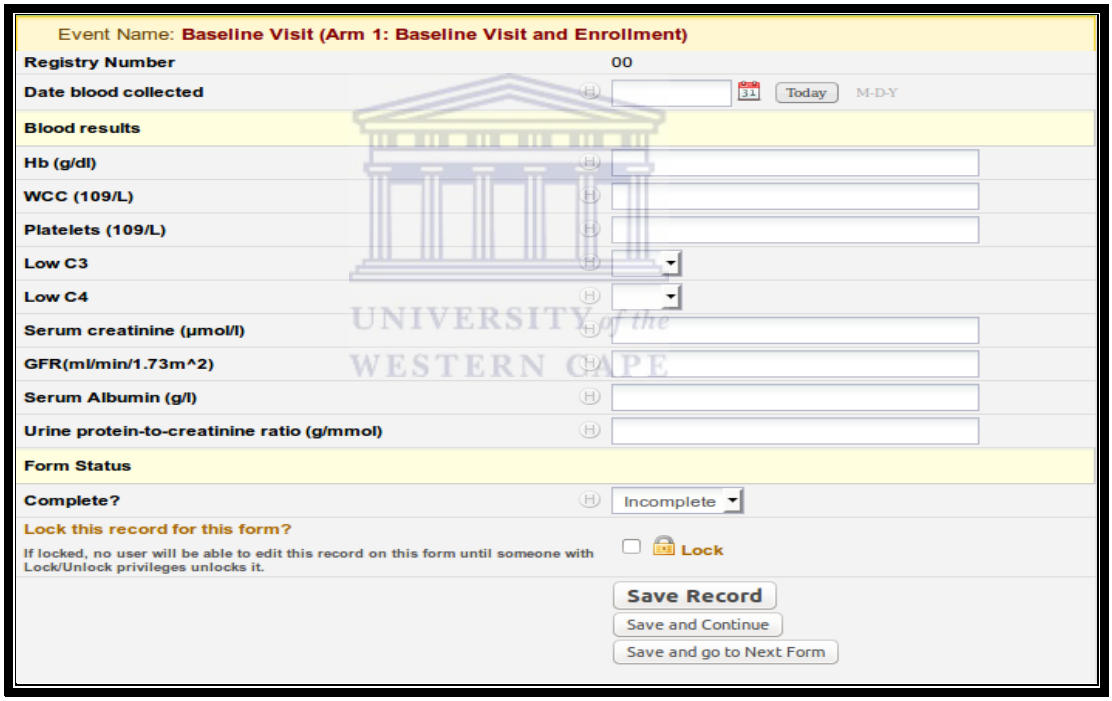


Figure 4.3 Sample data entry form. The registry number is a unique patient number assigned to each participant. This form is completed at a baseline visit (as highlighted in red). This form also allows incomplete information to be entered, saved and highlighted as "Incomplete" which allows the user to return to the form once the information is available.

4.2.2 Data sourcing

The manner in which data is collected, verified or validated will help shape their use in a database. The selection of data elements to capture in the clinical database requires balancing of factors such as the importance of a particular data item for the

integrity of the database and for the overall analysis of primary patient outcomes, the reliability of the collected data, and the incremental costs associated with their collection. Specific data elements are selected with consideration for established clinical data standards and common data definitions. It is important to determine which data elements are absolutely necessary and which are desirable but not essential. In choosing measurement scales for assessing patient-reported outcomes, it is preferable to use scales that have been appropriately validated, only when such tools exist.

Patients entered in the clinical database will be recruited from participating hospitals, based on the exclusion-inclusion clinical criteria that will be specified by participating clinicians. Since the clinical database is set-up as a longitudinal database, patients will be followed up for a period not less than five years and in some cases where possible for ten years or more. Detailed patient data will be collected at presentation and at annual reviews. Currently, the clinical database has been initiated for a project in the Western Cape, South Africa at Groote Schuur Hospital where 250 patients have been enrolled since 2012 and are currently being entered into the database. Other centers in South Africa as well as in Nigeria, Ghana, Senegal, Morocco, Guinea and Kenya have indicated interest in joining the clinical database. Since the database is going to capture patient data, ethics approval will be needed for each participating center, according to their rules and regulations. Therefore, signed consent will be obtained from all participants before entry into the database. Importantly, patient confidentiality will be maintained at all times throughout the entire period that clinical data will be kept in the database.

As the database goes live, it will be very useful to collect some metrics. For instance, metrics should be collected on how long it takes to complete a CRF, for example, patient demographics. The time taken to complete the CRF will be correlated with the accuracy of the information that is paper based in the clinical notes. Also, other clinical researcher will be invited to go through the database to evaluate the usefulness of the database in disease cohort standardisation, characterisation and potential for scaling up.

4.3 Results

4.3.1 Database home page

The home page provides a brief introduction and overview of the REDCap system (Figure 4.4). It also shows different tabs that one can use to access the database, the control center which is only accessed by a designated systems administrator and a tab that is used to create different projects (Figure 4.4). Importantly, the home page shows a link that can be used to create a new database (Figure 4.4).

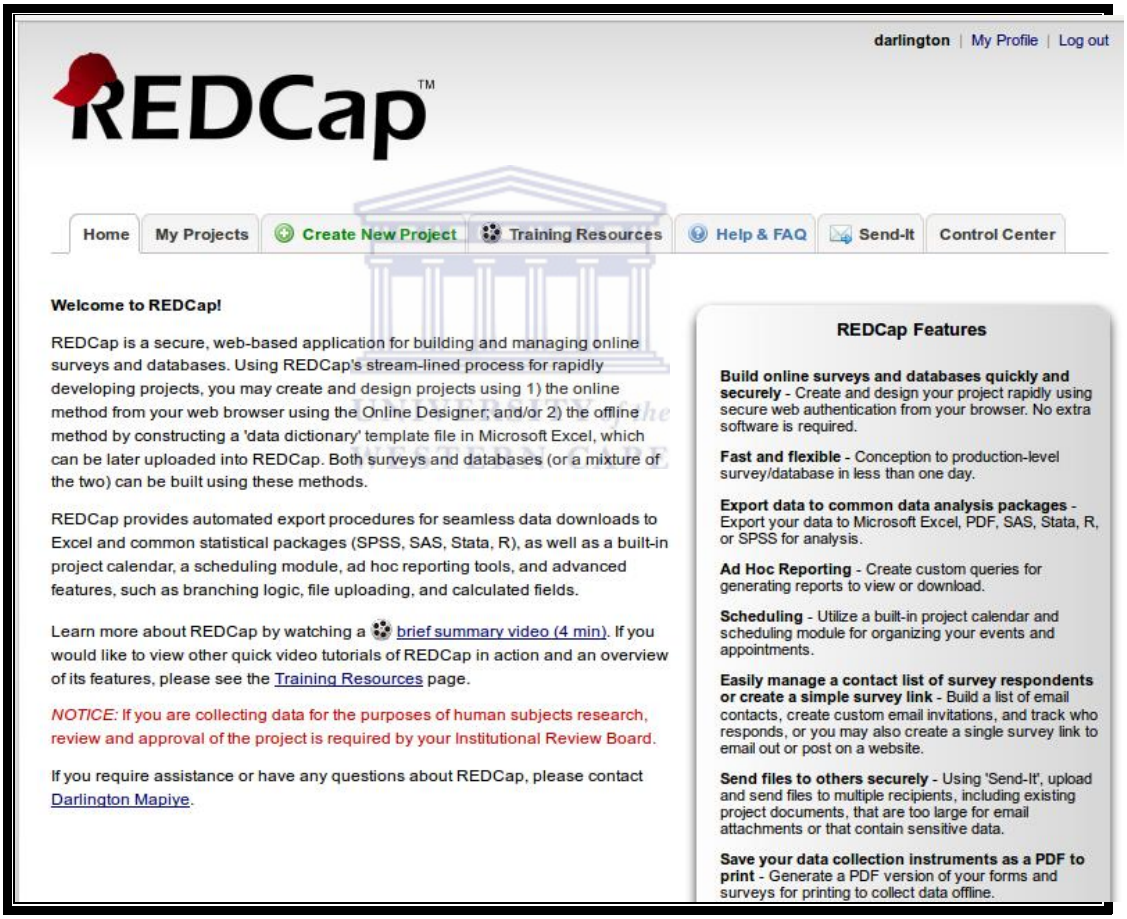


Figure 4.4 Clinical database home page. Once access to REDCap is granted this is the first page accessible to the user. Depending on the assigned privileges some of the tabs such as the "control center" and "Create New Project" will not appear on this page. The "My Projects" tab will contain a list of databases accessible to the user.

4.3.2 Database access

The clinical database is accessed through "My projects tab" (Figure 4.5). Although it

possible to have many databases set-up under “My projects” users will only access the database for which access privileges have been assigned and they will not be able to view any other databases.

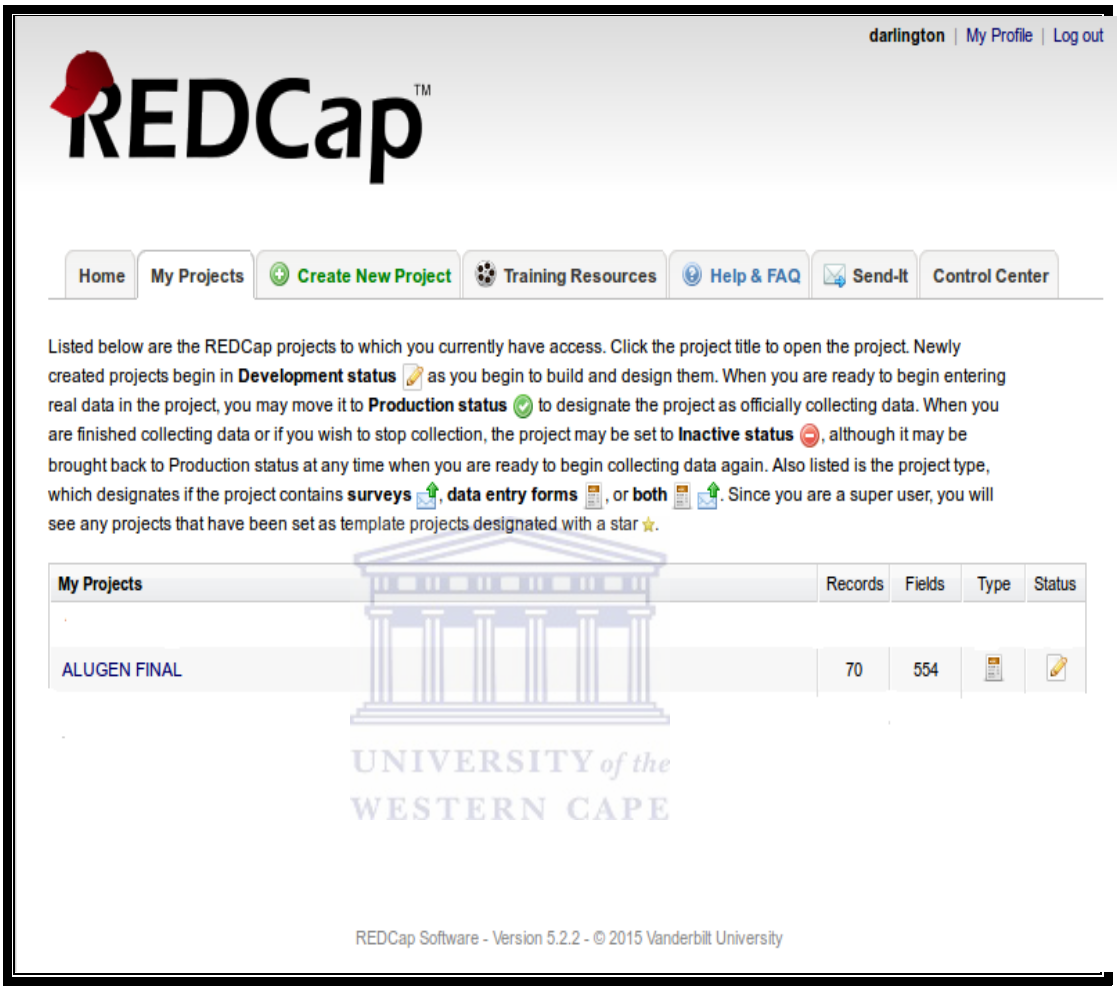


Figure 4.5 Database access. In blue is the ALUGEN FINAL clinical database that has been created and already contains 70 patient records.

4.3.3 Database functions

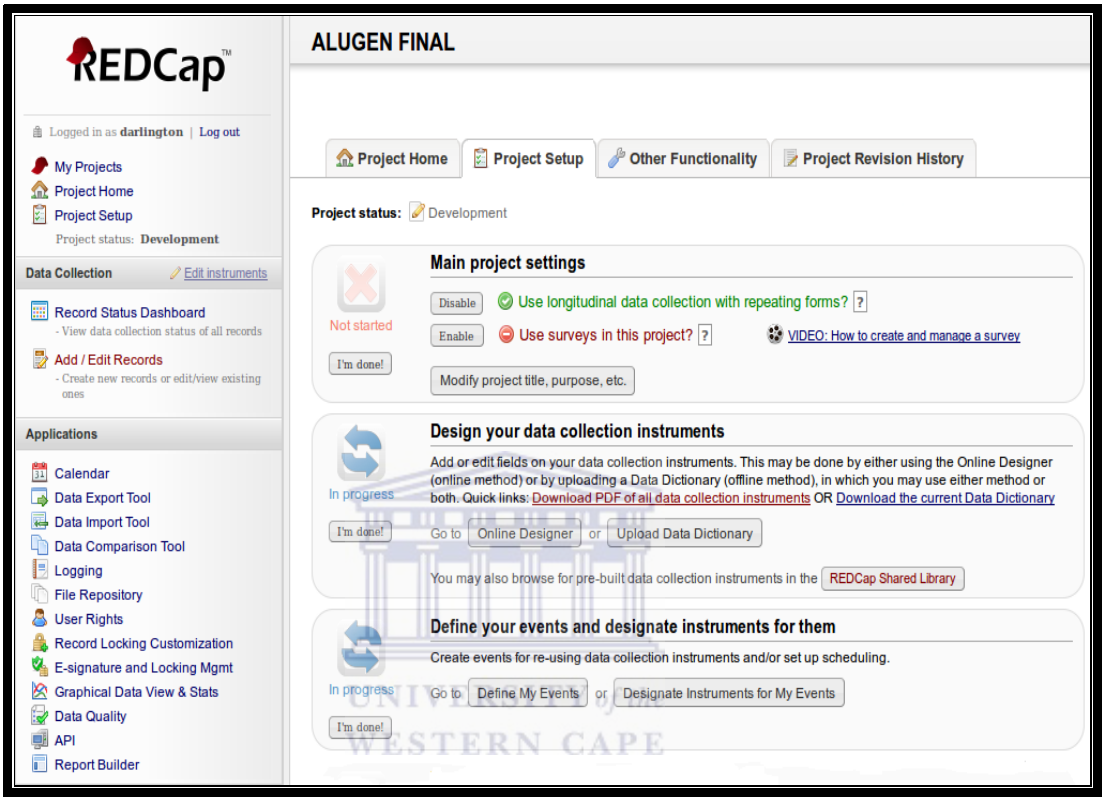


Figure 4.6 Database functionality. Once logged into the database the user will access the "Project Home" tab where data entry can begin. The data collection dash board and the applications tabs will also be accessible to the user and these are useful during data management.

Once a user has logged into the clinical database, they will access the "Project Home" tab where they can begin to enter patient data (Figure 4.6). The clinical database is designed with two different arms the "Baseline arm" and the "follow-up arm". The user will then chose the correct arm according to the patient data they want to enter. Data collection forms can easily be accessed through the "data collection" dash board (Figure 4.6). Also, the user will have access to "Applications" dashboard, which is useful for data management as the user can generate simple reports, store clinical protocol documents and produce a few statistics. The applications are, however, not designed to perform detailed statistical analysis. This is, however, possible by exporting the dataset in the format required for the statistical package of choice.

4.3.4 Real time data entry

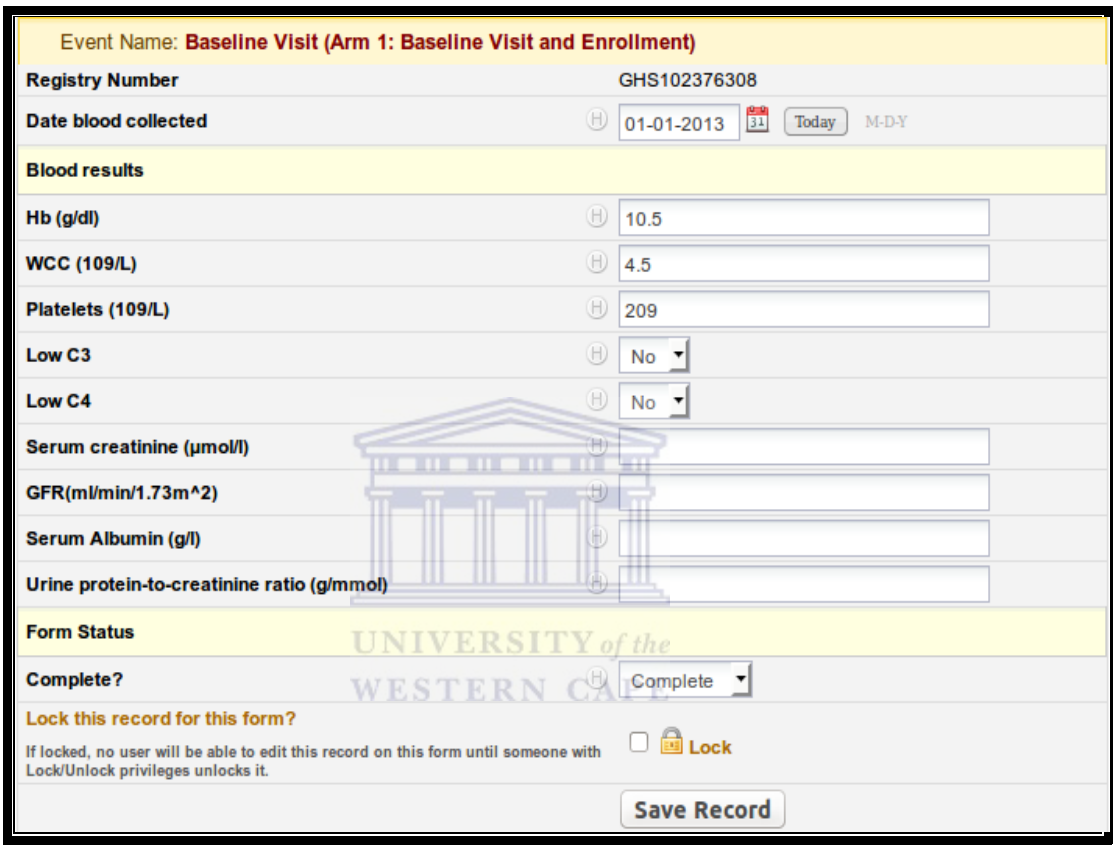
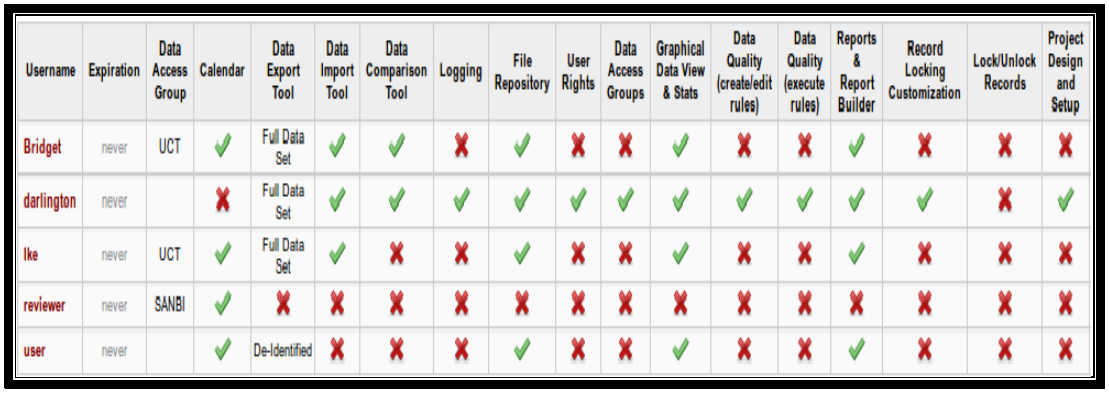


Figure 4.7 Sample completed data entry form. The data entry form was completed for a baseline patient visit. Data entry was performed using an intuitive user interface.

Once data entry has been completed the form is marked as "complete" (Figure 4.7). Important records can also be locked and be accessible only to those people with high level security privileges. This is important as sensitive clinical information can be kept from being edited by unauthorised personnel. Such security assignments are detailed in Figure 4.8.



Username	Expiration	Data Access Group	Calendar	Data Export Tool	Data Import Tool	Data Comparison Tool	Logging	File Repository	User Rights	Data Access Groups	Graphical Data View & Stats	Data Quality (create/edit rules)	Data Quality (execute rules)	Reports & Report Builder	Record Locking Customization	Lock/Unlock Records	Project Design and Setup
Bridget	never	UCT	✓	Full Data Set	✓	✓	✗	✓	✗	✗	✓	✗	✗	✓	✗	✗	✗
darlington	never		✗	Full Data Set	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓
Ike	never	UCT	✓	Full Data Set	✓	✗	✗	✓	✗	✗	✓	✗	✗	✓	✗	✗	✗
reviewer	never	SANBI	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
user	never		✓	De-Identified	✗	✗	✗	✓	✗	✗	✓	✗	✗	✓	✗	✗	✗

Figure 4.8 Database comprehensive user rights assignment. User privileges are assigned by a designated database administrator. Green ticks highlights functions for which a user has been granted access while the Red Cross shows utilities that a user will not be able to access.

4.4 Discussion

In studies of CKD and its risk factors (e.g. SLE), clinical databases can be a powerful tool that may be utilised to observe the course of a disease, to understand variations in treatment and outcomes, to examine factors that influence prognosis and quality of life, to assess level of patient care including effectiveness, appropriateness, disparities and diversity of patient therapies (Lu et al., 2010; Sam Lim et al., 2009; Villa-Blanco and Calvo-Alén, 2012). Also, data from clinical databases can be of great use to analyze the magnitude and distribution of disease manifestations thus, increasing knowledge on the burden of a disease which can contribute significantly to improving health planning (Sam Lim et al., 2009).

The paucity of prospective longitudinal cohort studies of SLE patients in SSA, and the difficult in diagnosing and managing the disease makes multi-center clinical databasing imperative (Tiffin et al., 2013). Clinical research in SLE at individual centers is complicated by the difficulty of accruing enough patient numbers (Tiffin et al., 2013). In this context, the development of a multi-center clinical database will allow the recruitment and pulling together of patient cohorts from different geographical locations and the collection of standardised, compatible datasets which has the potential to increase our knowledge regarding the clinical course and management of the disease (Villa-Blanco and Calvo-Alén, 2012). By including patients from different ethnic and geographic backgrounds investigators are likely to enhance our understanding of inter-ethnic and regional variations in disease

expression, and will be able to explore the role of genetic factors in predisposition and disease expression (Sam Lim et al., 2009).

Furthermore, identification and analysis of patients with SLE at different participating sites may assist to estimate incidence and prevalence rates which will greatly improve our understanding of SLE, its public health burden, and implications for health care planning. The large number of SLE patient data to be collected and stored in the database will allow for greater power to look at differences in disease presentation, progression and outcome across different sex, ethnicity and age groups (childhood onset versus adult onset) of patients in different geographical areas (Lu et al., 2010; Sam Lim et al., 2009; Villa-Blanco and Calvo-Alén, 2012). Also, a multi-center clinical database that pulls together patient data from different geographical locations can be utilized to develop statistical models to predict outcome based on prognostic and treatment factors that may be appropriate for resource limited areas such as SSA. Going forward this multi-center clinical database will enable participating centers to formulate and answer their own research questions, expand African research infrastructures and increase support for upcoming researchers. Importantly, it will also provide a clearly well-defined and phenotypically well characterized cohort which can be used to undertake future large scale genomic studies of SLE in Africa. The extension of use of this database to other African researchers and clinicians working in SLE will increase collaboration and an integrated pan-African initiative towards addressing SLE in African patients.

4.5 Limitations

Despite rigorous efforts to minimize poor quality of data entered, the clinical database has some other limitations. The database can be affected by selection bias related to the method of choice for participating centers. Another limitation emanates from differences in referral to identify potential patients depending on various factors such as the level of care or the presence of certain organ-specific manifestations that encourage referral to other specialties. Also, incomplete follow-up data which may be caused by the delay in disease diagnosis is another limitation. For complex, non-communicable diseases such as SLE, some relevant patients may fail to be captured.

To mitigate this risk, a comprehensive, active surveillance system with numerous patient-finding sources needs to be designed. Also, patients may migrate in and out of the catchment area for medical care or residency and this makes it difficult for them to be followed up.

4.6 Conclusion

Clinical databases can be used for much more than estimating incidence and prevalence of a disease. They also allow a cross-sectional assessment of the association of a variety of factors, including socioeconomic factors and patient outcomes not systematically incorporated in other studies. Once completed, the clinical database will provide a well-defined cohort that will be prospectively followed over time to address important issues with respect to disease progression and management. Clinical databases can be integrated with electronic health records (EHRs) to directly support evaluation of care delivery and patient outcomes and also broaden knowledge of clinical service patterns and processes. Although the inferences that can be drawn from observational data may be limited by selection bias, clinical databases are valuable tools in planning clinical research. In an era where much research funding is directed at hypothesis-driven research, the importance of clinical databases in developing clinical research hypotheses should be seriously considered.

5 Summary of key findings and future direction

The human genome comprises 3 billion base pairs. Approximately 85% of disease-causing mutations have been found in the exonic regions. To date, there are approximately 180,000 known exons, which constitute about 1.5% of the human genome, or approximately 30 million base pairs. Mutations in the exonic sequences are predicted to be more likely to have severe consequences than those in the non-coding regions of the genome (Choi et al., 2009b; Ng et al., 2009a). Therefore, the goal of sequencing the exome is to identify genetic variation responsible for human diseases without incurring the high costs associated with whole-genome sequencing (Ng et al., 2009a).

Advances in sequencing technologies are making previously intractable genetic analyses now possible (Choi et al., 2009b; Ng et al., 2009a). Utilizing such technologies to make precise genetic diagnoses may not only help distinguish between diseases with related phenotypic and histopathologic patterns, but also permit researchers to draw conclusions from analyses performed on a small number of individuals within a single family. Technologies such as massively parallel DNA sequencing can increase the affordability, efficiency, accuracy, and speed of diagnosis (Choi et al., 2009b; Ng et al., 2009a). These technologies are becoming more readily available to assist in the accurate diagnosis of many genetic disorders including genetic disorders of the kidney (Malone et al., 2014).

Many disorders of the kidney can present as unclear collections of overlapping and non-specific phenotypes (Malone et al., 2014). Genetic analysis of disease-causing variation in disorders that are caused by single-gene defects, as is the case in many kidney disorders, is the most robust diagnostic approach for accurate diagnosis (Edwards et al., 2014; Malone et al., 2014; Xiu et al., 2014). In this work I describe a South African family in which individuals presented with ESRD, although clinical and histological presentations were not helpful in elucidating the cause of the disease. Therefore, I considered alternative diagnostic methods that are based on performing genomic evaluation to explore the primary causes of the disease in this family. In

families where there is evidence of a familial inheritance pattern, genetic analysis may provide the simplest and most efficient method for making an accurate etiological diagnosis. In this thesis, emphasis is put on the importance of the use of next generation sequencing technologies on samples from patients and their families, particularly when meticulous clinical and histopathological information is lacking. Going forward high-throughput genetic tools are becoming increasingly important diagnostic methods.

5.1 Major contributions of this work

Apart from unpacking plausible genetic mechanisms underlying a rare and atypical familial kidney disease of unknown aetiology in an African family using high-throughput sequencing technology, this thesis also provides a robust and well-designed, reusable computational pipeline for analysis of human exonic data. Importantly, the project addresses in detail a current active area of research which deals with annotation of human genomic variation and candidate gene identification. Therefore, this work contributes significantly to the literature on gene discovery for complex renal phenotypes and computational analysis of composite high throughput genetic data.

In addition, the thesis also proffers a solution to the problem that is currently faced by many clinicians in sub-Saharan Africa, of lacking standardised methodologies for collection of clinical data. Collection of standardised patient data is a very crucial first step towards advancement of genomic studies in Africa especially in the era of declining sequencing cost, which allows research to sequence the entire human genome and begin to unravel genetic variation that is associated with, causes and predisposes people to certain diseases. The collection of standardised clinical data will become increasingly important as researchers and clinicians venture into the area of precision medicine and also try to understand and integrate additional data about environmental factors that are associated with diseases. Briefly detailed below is an outline of the major contributions of this thesis.

5.1.1 Clinical databasing

Lack of phenotypically well characterised disease cohorts is one of the major things that hamper the successful undertaking of genomic research in resource limited countries. To help alleviate this caveat clinical databasing once implemented effectively and efficiently will enable the collection and storage of well-structured clinical data. Given that genomic studies of rare complex diseases often fail because of lack of or limited availability of patient samples, multicentre clinical registries would enable pooling together of samples from different researcher to increase sample sizes, conduct research to understand the environmental factors that are associated with diseases in different geographical areas and in the process increase the possibility of scientists collaborating, sharing expertise and increasing research capacity. Importantly, a clinical database would help to identify and stratify appropriate individuals for future genetic/genomic studies based on a full and standardised phenotype. The individuals enrolled will be from different geographic locations and diverse bare grounds establishing a basis to also study the effect of environmental factors on Lupus. The clinical database has an underlying SQL relational database that could be joined to a genomic/genetic database through identifiers such as a patient hospital number. A graphics user interface can then be built to query both database and leverage the use of integrated clinical and genomic data for a patient, providing a holistic view of a patient. This will become more important as clinical medicine moves towards precision patient care with effective treatment plans being designed for each patient based on their genetic profile and clinical parameters, similar to the approach used by BioMart.

Once properly implemented clinical databases can be used to evaluate outcomes ranging from disease presentation, to prognosis, to the safety of drugs and effectiveness of therapies (Singh et al., 2012). Also, epidemiological research on disease occurrence and distribution, disease risk or etiology and disease prevention can be done using data from clinical databases (Sam Lim et al., 2009). Furthermore, data from clinical registries can be compiled for the purpose of health care planning, implementation and evaluation in a well-defined population. Importantly, data from clinical databases can be used in African genomic studies for generating research

hypothesis that maybe undertaken to better understand the causes, prognosis, management and outcomes of diseases.

5.1.2 Analysis of exome sequencing data based on African samples

To date and to our knowledge, this is the first study that sought to understand the genetics of rare familial clustered end stage renal disease in a South African family, where primary clinical diagnosis and histological analysis lacked enough information to provide a conclusive cause of the disease. Thus, the genetics basis of the disease in this family was investigated. Given the genetic diversity in the Africa population, this study contributes significantly to understanding genetic mechanisms that may underlie ESRD in African populations, especially considering that most studies that seek to unpack the genetics of complex familial clustered renal phenotypes have been undertaken in other populations. By understanding the genetic mechanism underlying this rare and atypical ESRD phenotype in African patients, we gain more general insights into mechanisms of dysregulation of kidney function that may also be relevant for other forms of the ESRD.

The pipeline implemented for this project utilised joint variant-calling (JVC). JVC has two major advantages. Firstly, the JVC algorithm considers the alignments of all samples simultaneously to estimate the probability that a given locus is variable in the population, resulting in more accurate variant calls for each individual sample considered. Second, JVC such as GATK Haplotype caller provides missing genotypes. Most variant callers by default will not produce a variant call for missing genotypes. Thus, homozygous reference sites are indistinguishable from sites where no genotype information is available, for example, due to poor sequence quality or low depth of coverage. When both affected and unaffected family members are processed through JVC algorithms, all variant sites in all samples are consistently called for missing genotypes.

In addition, a reusable computational pipeline for analysis of exome sequencing data was designed. The pipeline also addressed a key issue of high levels of genotyping errors that are associated with whole exome sequencing analysis. A probabilistic

recalibration model, which calibrated SNPs and Indels separately, was implemented to improve the quality of genotype calls. The algorithm used was developed to create a model of true-positive variants trained on accurate variant calls using for example, HapMap and other highly validated variant sites. This was followed by stringent filtering parameters to improve further the accuracy of genotypes. Currently, the pipeline is being implemented by other researchers within the institute to analyze exome sequencing data. Also, the analysis pipeline incorporated structural variation analysis of exome reads, this included copy number variants (CNV) and variant number tandem repeats (VNTR), which is a current area of active research.

In this work, I have also demonstrated that it is indeed possible to sequence only a few family members where the cause of the disease is suspected to have a significant genetic basis and still be able to uncover the plausible genetic mechanisms driving the disease in the family.

5.1.3 Quality control of exome sequencing data using relatedness testing

In genetic analyses, knowledge of relatedness may be used to estimate genetic parameters such as heritability and genetic correlations. When a family pedigree is known, genetic relatedness between individuals can be calculated from the pedigree and can be used to estimate how much DNA is shared amongst family members and infer if this is consistent with the relationships described in the family pedigree. This is particularly important when one is investigating the plausible genetic cause of a familial disease and "TRUE" paternity and relations are sought. In this work, family relatedness using exome sequencing data were performed using a linear mixed model approach implemented in PLINK. In each case I used a full set of SNPs to perform the relatedness calculations, using VCFtools and PLINK.

Also, using relatedness analysis in order to ensure that samples were related as reported, I was able to discover a sample aliquot that was duplicated and sequenced as two separate samples. For instance, in the initial relatedness analysis 999 (control sample-Aunt) and 888 (affected Nephew) were identified to share approximately 99% of their DNA. This raised a flag as an Aunt and Nephew are expected to share

approximately 30% of their DNA. Further investigation revealed that 999 had been sequenced twice, once as the Aunt and the other as the Nephew. These computational results were confirmed by PCR sex check on the samples. At this point, had computational relatedness analysis not been done I might not have identified this anomaly. All samples were then double-checked and where necessary sequenced again. This illustrates the importance of incorporating computational relatedness analysis as a step in the analysis of familial exome sequencing data, to ensure validity of findings.

5.1.4 Statistical probabilistic variant prioritization of exome sequencing data

A lot of systematic technical differences exist in variant calls identified between affected and unaffected samples. This can be a major source of false-positives when variants are prioritised in the analysis of exome sequencing data for familial studies. Thus, a number of steps were performed in order to improve variant prioritisation to reduce such technical artefacts.

A major weakness of many variant prioritization tools is that they can only prioritize variants within phylogenetically conserved coding regions. Thus, these algorithms have poor coverage across the genome. For example, SIFT, MUTATION tester and PolyPhen can score only 60% and 81% of the human proteome, respectively (Adzhubei et al., 2010b). Another weakness of these approaches is that they make no use of allele frequency information. It has been demonstrated that minor allele frequency (MAF) is negatively correlated with purifying selection pressure.

In this work, I have demonstrated that it is indeed possible to utilise probabilistic statistical models to prioritise variants using only exome sequencing data obtained from merely a few sequenced family members. The probabilistic model applied in this project combines both amino acid substitution (AAS) information with variant

frequency information to prioritise variants, allowing it to score all variants with more accuracy no matter where they lie in the genome. Also, the model makes use of missing genotype information, which substantially improves the signal-to-noise ratio in variant prioritisation. A lack of missing genotype data in affected or unaffected individuals can be a significant source of error for all downstream analyses and interpretations. Importantly the model also supports analysis of small insertion and deletion (indel) variants.

5.1.5 Multiple variants theory

Most Mendelian diseases are caused by a single highly penetrant rare or novel variant. The first successful study to unravel the genetic cause of a rare disease using NGS techniques implicated a single variant in a single gene (Choi et al., 2009b). Given this success most researchers that have sought to unravel the genetics underlying rare diseases using NGS methods in most cases have identified a single variant in a single gene (Ng et al., 2010a). This also has become the norm in most studies that have been published (Bilgüvar et al., 2010; Sankaran et al., 2012; Woo et al., 2013; Worthey et al., 2010).

In this work, however, I have observed that this may not necessarily be the case. For instance, a striking pattern where a number of novel and rare variants were identified in genes that are located on the same chromosome and very close to each other was observed. One hypothesis that can be postulated is that possibly the disease is not caused by a single defective gene but rather a combination of variants in different but closely located genes, whose combined effect may result in the observed disease phenotype. Further analysis would then be required to establish if these variants are in perfect linkage disequilibrium. Another way of looking at this pattern is that, maybe there is a combination or a single haplotype segregating with the disease in this particular family. These are all hypotheses that could potentially be supported by the results obtained in this work. Clearly, given the “Narrative potential of the human

genome" it can be seen that focusing solely on finding a single novel or rare variant as causal for a genetic disease may result in loss of valuable information that may in reality explain important disease mechanism and etiology.

Similarly, a novel and a rare variant were identified in the same gene. Though it is one gene that is affected, it maybe that only the presence of the two variants in the gene will result in the affected person having a severe form of the disease or that only when the two variants are present will an individual develop the phenotype. Therefore, it's not always the case that extreme phenotypes are caused by a single very rare or novel variant. It might be that the presence of both variants will trigger the disease but the inheritance pattern appears that of a Mendelian single mutation because the variants lie so closely together on the genome. Overall, the results obtained in this work have opened us to other ways of unpacking genetic bases of rare complex renal phenotypes and other idiopathic diseases in general.

5.1.6 Structural variation inference from exome reads

The first type of structural variation investigated in this work is the short tandem repeats (STR). STRs, also known as microsatellites, are a class of genetic variation with repetitive elements of 2–6 nucleotides that consist of approximately a quarter million loci in the human genome. STR expansions have been implicated in the aetiology of a number of genetic disorders, such as Huntingon's Disease and Fragile-X Syndrome. STR variations are, however, not routinely analysed in exome sequencing studies mainly due to a lack of adequate computational tools (Treangen and Salzberg 2011). STRs pose a significant challenge to high through put sequence analysis. First, not all reads that align to an STR locus are informative. Second, mainstream aligners, such as BWA, generally exhibit a trade-off between run time and tolerance to insertions/deletions (indels) (Li and Homer 2010). Thus, profiling STR variations even for an expansion of three repeats in a trinucleotide STR would require a cumbersome gapped alignment step and lengthy processing times. Despite these difficulties, STRs were profiled in all the exomes of the sequenced family members using LobSTR (Gymrek et al., 2012). The algorithm scans genomic libraries, flags informative reads that fully encompass STR loci, and characterizes their STR

sequence. This was a computationally intensive analysis that was optimised at different stages of the analysis. Despite none of the STRs segregating with disease status in this family, the analysis has demonstrated and illustrated an efficient bioinformatics process of computational profiling of STRs using only exome reads obtained from a few family members. Statistical and bioinformatics methods for inferring STRs are an active area of research currently. Also, the analysis highlighted the need for having enough data storage capacity in order to optimise the computational process.

The second type of structural variation investigated in this work is copy number variation (CNV). In contrast to whole-genome sequencing data, exome sequencing results in non-uniform read depth between captured regions and strong systematic biases between batches of samples sequenced. These biases make exome sequencing unsuitable for CNV detection algorithms. In this study, I combined read-depth data from exome sequencing with singular value decomposition (SVD) methods to discover rare CNVs that may segregate with disease status in this family. Therefore, I was able to demonstrate that it is indeed possible to use exome sequencing data from a few family members to scan the coding region of the genome for CNVs that may segregate with disease. This sets the platform for developing further computational algorithms for detecting CNVs from exome sequences.

5.1.7 Genetics underlying rare complex renal phenotypes

Even though routine analysis of urine samples can be helpful to indicate the origin of some kidney disorders, the assessment of kidney disease activity and progression is still mainly based on crude markers such as serum creatinine and haematuria/proteinuria. The descriptive assessment of kidney biopsy specimens with use of light and electron microscopy, supplemented by a small set of immunological marker proteins, is still the diagnostic gold standard. Accurate diagnosis of the primary cause of an individual's kidney disease is essential for proper management.

Many rare kidney diseases have a different prevalence in different populations and have substantial clinical heterogeneity in presence, age of onset, severity, and

progression of symptoms. Most of the studies undertaken to ascertain these have been performed in other populations outside of Africa. Therefore, different incidence rates in populations provide support to a role for genetics in understanding the pathogenesis of kidney disease especially in African populations where there are limited studies done. Over the years, an increasing number of rare kidney diseases that were previously considered to be single disorders have been shown to be aetiologically heterogeneous. What is usually observed is that different underlying genetic abnormalities can affect the same biological pathways and give rise to similar clinical, biochemical, and histopathological features. The imperfect diagnosis made by traditional methods is largely explained by their inability to elucidate underlying molecular disease mechanisms.

The advent of next-generation sequencing techniques has ushered in a new era of diagnostic capability that will improve diagnosis efficiency for genetic renal diseases through simultaneous investigation of all relevant genes for a given kidney phenotype at much reduced cost and turn-around time. In this work, I have successfully applied next-generation sequencing techniques to complement primary diagnostic methods in order to understand the cause of ESRD in a South African family. Dysregulated COL4a1 and COL16a1 proteins were implicated as potential causes of the disease in this particular family. Collagen genes are interesting candidates for ESRD, as a number of exome sequencing studies implicated these genes in rare complex renal phenotypes.

In a Chinese Hans family spanning over 5 generations, four patients presented with heterogeneous clinical phenotypes of glomerosclerosis, while none of them showed any clinical features of either sensorineural hearing loss or typical ocular abnormalities, clinical diagnosis failed to provide a conclusive cause of the disease in this family. Exome sequencing was undertaken and a novel COL4a5 mutation was identified (Xiu et al., 2014) (Xiu et al., 2014). Exome sequencing was undertaken in three families who presented with similar clinical features as the family we have studied where renal biopsies at that time were inconclusive. A novel variant in COL3A3 gene and a missense mutation COL4A3 were implicated (Lin et al., 2014). Similarly, in the South African family I have also identified rare variants in Collagen genes. In another study a girl aged 6 presented with haematuria and her sister aged 5

presented with haematuria and proteinuria. Family history showed multiple individuals suffering from end stage renal failure from the paternal side of the pedigree. Exome sequencing was undertaken, a mutation in COL4A5, a gene known to cause Alport syndrome was identified (Gibson et al., 2013). This led to the diagnosis of the girls being resolved to Alport syndrome. COL4A3 and COL4A4 variants were identified using exome sequencing in cohort of 70 families with complex renal phenotypes (Malone et al., 2014).

As illustrated above, this study contributes significantly to the growing spectrum of the "*collagen genes"* and their potential role in rare complex renal phenotypes. This is important as comparisons can now be drawn regarding the genetics of kidney diseases in African populations with other populations, as this study is based on an African Family. Thus, clearly we have contributed significant to the overall knowledge base of kidney diseases by showing that some African renal diseases follow similar disease mechanisms to those in well studied populations and that some rare renal diseases have overlapping histopathological features despite being caused by defects in different genes.

5.2 Concluding remarks

Beyond sequencing disease specific gene panels, exome sequencing will soon become part of routine molecular diagnostics, improving further disease diagnostics. Sequencing based technologies are also increasingly being applied to individual cells, with the aim to integrate genomics, transcriptomics, epigenomics, and proteomics for multilevel analysis of cellular mechanisms. These analyses will need robust single-cell isolation, a potentially challenging task for a heterogeneous tissue such as kidney. In this work, WES identified 3 pathogenic variants in COL4A1, COL16A1 and ICAM1 in 5 African family members with previously unexplained inherited kidney disease. These findings highlight the clinical range of collagen related nephropathies and will help resolve diagnostic confusion arising from incomplete clinical and histological findings, allowing appropriate counselling and treatment advice to be given. Despite progress in understanding of molecular causes of rare inherited kidney diseases, the pathways for most inherited nephropathies still need to be explored. Poor

appreciation of genetic studies by health-care providers is of concern. Even for well-defined disorders the use of genetic testing remains rare, mainly because of high cost and long turnaround times for conventional genetic screening, the preconception that a genetic diagnosis will not affect clinical management, insufficient genetic literacy, and differences in access to genetic tests.

Clinical databasing will go a long way to enable clinicians to collect and store standardised clinical data for their patients. This will allow accurate phenotyping to be done, which is a key necessity for undertaking successful genetics analysis. Providing this important resource for clinicians creates an important platform for genetic diagnostics to be used effectively and implemented in resource limited countries as an important part of disease diagnosis where primary diagnosis lacks useful information to aid clinical management of diseases. Therefore, limitations notwithstanding, this work addressed in detail the following:

(a) The literature on kidney disease in African population has been reviewed intensively and clearly highlights the gap that exists between Africa and other developing countries in tackling the scourge of non-communicable disease like ESRD.

(b) The problem of collecting standardised clinical data that is crucial for carrying out genetic analysis based on Africa populations has been addressed. A database was designed and this work has been accepted for publication in Lupus (Hodkinson et al., 2015). This database will be the first Pan-African database intended for the collection of standardised patient data across different Africa countries. This will allow examination of hypothesis concerning disease genetics, aetiology and health outcomes of patients.

(c) I have demonstrated that it is indeed possible to undertake a high throughput genomic study to investigate causes of disease in an African family. To my knowledge, this is the first study performed to understand the genetics underlying familial clustered ESRD in an African family.

(d) A clearly designed pipeline for analysis of exome sequencing data was designed and implemented in this study. The pipeline also included probabilistic statistical models to help analyse the data.

(e) Novel and rare genetic variants underlying ESRD in this South African family were identified. This adds to the pool of variants that have been implicated in patients with familial clustered ESRD.

(f) Pathway and functional analysis identified cellular and molecular regulatory mechanisms that are related to kidney disease, in which prioritised genes were enriched.

(g) I have shown that is possible to computationally infer structure variation such as Copy Number Variation and Short Tandem Repeats using exome sequencing data from a few family members.

(h) Importantly, this work has contributed significantly to the wider spectrum of collagen genes and their potential involvement in rare complex renal phenotypes. This adds more evidence to the crucial role that these structural proteins may have in the pathogenesis of ESRD.

5.3 Future direction

The abundance of genetic and molecular information generated by next-generation sequencing poses a new challenge because bioinformatics capacities and analysis methods need development. The characterisation of candidate disease genes and individual mutations needs to be studied further. In this family no formal testing was undertaken to establish clinical features of either sensorineural hearing loss or related ocular abnormalities something that is going to be pursued with the clinicians. Collagen genes implicated in this study have also been implicated in patients with Alport syndrome, hearing loss and ocular abnormalities are some of the key symptoms of this disease. At least two renal biopsies need to be performed in order to determine the primary cells of the kidneys that are affected. This is important as one may then investigate the expression levels of implicated genes in these cells and begin to unpack the probable disease mechanism. Also, information from the renal biopsy combined with functional analysis of implicated genes might help explain a plausible disease mechanism in this family. Variants identified in this family will be genotyped in more unaffected people to investigate their frequency in the general South African population, to confirm whether they are truly rare alleles or whether they are present more frequently specifically in Africans/South Africans. These results will be further validated using Sanger sequencing. Going forward instead of sequencing the entire

exome for the family members a quick PCR can be performed to identify these mutations. Importantly, this can be done over several generations in this family and the variants identified by WES can then be concluded as truly causative for this family. This is one of the biggest advantages gained by undertaking WES in familial study and that opportunity exists to extend the results to other family members with minimum cost.

In light of the numbers of exomes anticipated to be sequenced and analysed in the near future, I believe that computational methods developed in this thesis will have widespread application for the discovery of both rare and novel single nucleotide polymorphisms as well as copy number variation and short tandem repeats in disease.

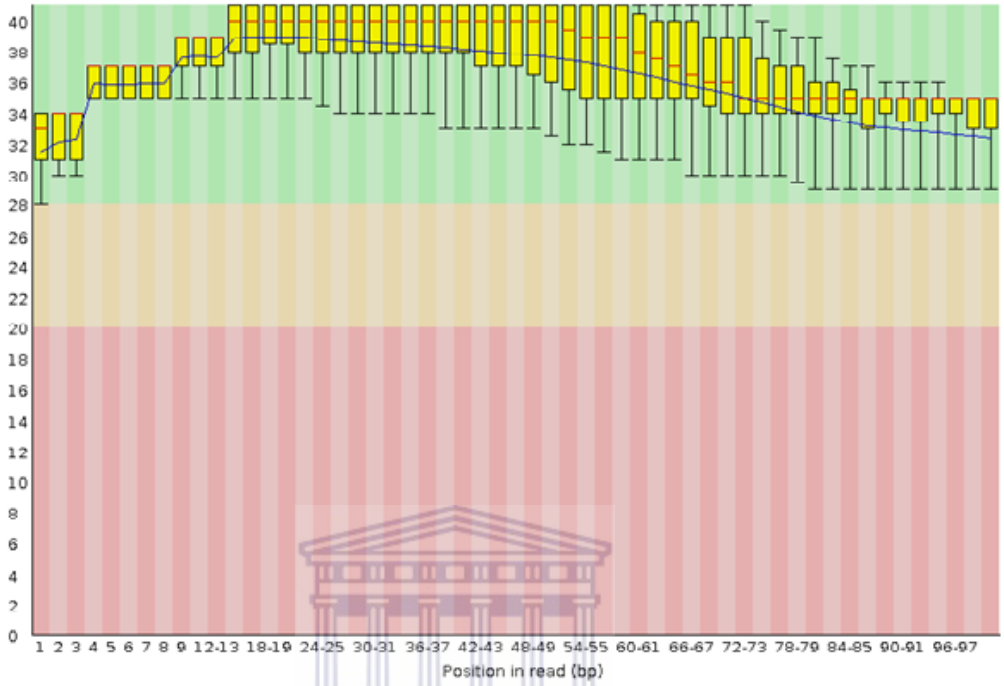
Increasing access to internet, computational facilities, and genetic analysis means that more clinicians can collect data for African diseases and the clinical database is a great prototype to help us, going forward, to understand effective ways to assist with clinical research databasing on the continent. Effective clinical databasing as I have demonstrated with nephrologists dealing with lupus nephritis patients is a first step towards improving awareness, quality and quantity of patient clinical data that clinicians within Africa capture. Going forward such an invaluable clinical research resource will become important in African genomic studies for both hypothesis-led and hypothesis-generating research approaches that maybe undertaken to better understand the causes, prognosis, management and outcomes of diseases. The clinical database I have designed has already raised interest as a blueprint for a similar type of Pan-African database for sickle cell anaemia.

UNIVERSITY of the
WESTERN CAPE

Appendix A. **Sample quality control information.** The quality control was done at the core sequencing facility before sequencing was done. All samples passed the accessed quality control steps. Library preparation was done using Agilent SureSelect Human All Exon Version 4. Sequencing was performed using the Illumina hiseq2000 machine.

Position in Gel	Sample name	Nanodrop Measurement (ng/µl)	OD 260/280	Vol. Loaded (µl)	Mass (µg)	QC results
1	222	70.2	1.86	2	1.4	Pass
2	555	182.3	1.87	2	3.6	Pass
3	666	102.2	1.82	2	2	Pass
4	777	144.1	1.84	2	2.8	Pass
5	888	192	1.86	2	3.8	Pass
6	999	109.2	1.86	2	2.1	Pass

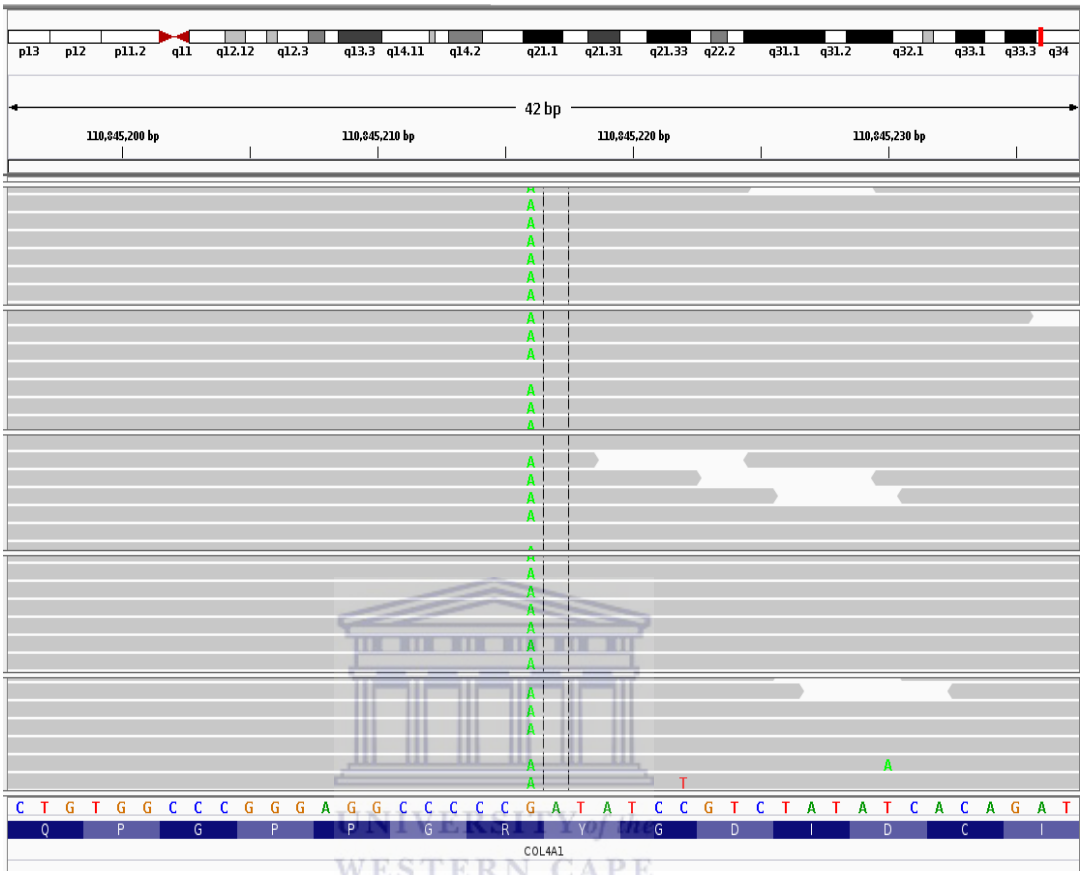
Appendix B. **FASTQ results for the unaffected family member.** Fhred score quality control scores where reported for each base position in the 100base paired reads. The quality scores were reported for paired end reads. The analysis was performed using FASTQC.



Appendix C. **Parameters for variant filtration using recalibration model.** The parameters were used to calibrate variants to ensure that only variants of high quality are retained for further analysis.

--filter Expression "MQ0 >= 4 && ((MQ0 / (1.0 * DP)) > 0.1)"
--filter Name "HARD_TO_VALIDATE"
--filter Expression "DP < 5"
--filter Name "Low Coverage"
--filter Expression "QUAL < 30.0"
--filter Name "Very Low Qual"
--filter Expression "QUAL > 30.0 && QUAL < 50.0"
--filter Name "Low Qual"
--filter Expression "QD < 1.5"
--filter Name "Low QD"
--filter Expression "FS > 150.0"
--filter Name "Strand Bias"

Appendix D. **COL16A1 [*p.T116M*] variant visualisation using IGV**. The variant is shown in green and it is present in all 5 affected family members while absent in the unaffected family member.



UNIVERSITY of the
WESTERN CAPE

Appendix E. 3D protein structure for ICAM1 and the identified variant. The modelled protein structure shows the location of the variant on the protein. The position of the variant on the protein is shown in red.

6 REFERENCES

Abboud, O.L., Osman, E.M., Musa, A.R., 1989. The aetiology of chronic renal failure in adult Sudanese patients. Ann. Trop. Med. Parasitol. 83, 411–414.

Abu-Aisha, H., Elamin, S., 2010. Peritoneal dialysis in Africa. Perit. Dial. Int. 30, 23–28.

Adam, J., Connor, T.M., Wood, K., Lewis, D., Naik, R., Gale, D.P., Sayer, J.A., 2013. Genetic testing can resolve diagnostic confusion in Alport syndrome. Clin. Kidney J. sft144.

Adzhubei, I., Jordan, D.M., Sunyaev, S.R., 2013. Predicting functional effect of human missense mutations using PolyPhen-2. Curr. Protoc. Hum. Genet. Editor. Board Jonathan Haines Al Chapter 7, Unit7.20. doi:10.1002/0471142905.hg0720s76

Agnes, H., Kalman, P., Jozsef, A., Henrik, B., Mucsi, I., Kamata, K., Sano, T., Naito, S., Okamoto, T., Okina, C., others, 2012. Clinical Nephrology-Epidemiology II. Nephrol. Dial. Transplant. 27, ii378–ii399.

Agrawal, S., Agarwal, S.S., Naik, S., 2010. Genetic contribution and associated pathophysiology in end-stage renal disease. Appl. Clin. Genet. 3, 65.

Akinsola, A., Adelekun, T.A., Arogundade, F.A., Sanusi, A.A., 2004. Magnitude of the problem of CRF in Nigerians. Afr J Nephrol 8, 24–26.

Al-Bhalal, L., Akhtar, M., 2005. Molecular basis of autosomal dominant polycystic kidney disease. Adv. Anat. Pathol. 12, 126–133.

Amos, A.F., McCarty, D.J., Zimmet, P., 1997. The rising global burden of diabetes and its complications: estimates and projections to the year 2010. Diabet. Med. 14, S7–S85.

Anantharaman, P., Schmidt, R.J., 2007. Sexual function in chronic kidney disease. Adv. Chronic Kidney Dis. 14, 119–125.

Andrew, S.E., Goldberg, Y.P., Kremer, B., Telenius, H. akan, Theilmann, J., Adam, S., Starr, E., Squitieri, F., Lin, B., Kalchman, M.A., others, 1993. The relationship between trinucleotide (CAG) repeat length and clinical features of Huntington's disease. Nat. Genet. 4, 398–403.

Ansorge, W.J., 2009. Next-generation DNA sequencing techniques. New Biotechnol. 25, 195–203.

Arar, N.H., Voruganti, V.S., Nath, S.D., Thameem, F., Bauer, R., Cole, S.A., Blangero, J., MacCluer, J.W., Comuzzie, A.G., Abboud, H.E., 2008. A genome-wide search for linkage to chronic kidney disease in a community-based sample: the SAFHS. Nephrol. Dial. Transplant. 23, 3184–3191.

Arogundade, F.A., Sanusi, A.A., Hassan, M.O., Akinsola, A., 2011. The pattern, clinical characteristics and outcome of ESRD in Ile-Ife, Nigeria: Is there a change in trend? Afr. Health Sci. 11, 594–601.

Austin, E.D., Ma, L., LeDuc, C., Rosenzweig, E.B., Borczuk, A., Phillips, J.A., Palomero, T., Sumazin, P., Kim, H.R., Talati, M.H., others, 2012. Whole exome sequencing to identify a novel gene (caveolin-1) associated with human pulmonary arterial hypertension. Circ. Cardiovasc. Genet. 5, 336–343.

Bamgboye, E.L., 2005. End-stage renal disease in sub-Saharan Africa. Ethn. Dis. 16, S2–5.

Bamshad, M.J., Shendure, J.A., Valle, D., Hamosh, A., Lupski, J.R., Gibbs, R.A., Boerwinkle, E., Lifton, R.P., Gerstein, M., Gunel, M., others, 2012. The Centers for Mendelian Genomics: A new large-scale initiative to identify the genes underlying rare Mendelian conditions. Am. J. Med. Genet. A. 158, 1523–1525.

Bansal, V., 2010. A statistical method for the detection of variants from next-generation resequencing of DNA pools. Bioinformatics 26, i318–i324.

Barsoum, R.S., 2005. Epidemiology of ESRD: a world-wide perspective. Kidney Dis. Dev. World Ethn. Minor. Lond. Taylor Francis 1–13.

Beaglehole, R., Yach, D., 2003. Globalisation and the prevention and control of non-communicable disease: the neglected chronic diseases of adults. The Lancet 362, 903–908.

Belinky, F., Nativ, N., Stelzer, G., Zimmerman, S., Iny Stein, T., Safran, M., Lancet, D., 2015. PathCards: multi-source consolidation of human biological pathways. Database 2015, bav006–bav006. doi:10.1093/database/bav006

Bell, C.J., Dinwiddie, D.L., Miller, N.A., Hateley, S.L., Ganusova, E.E., Mudge, J., Langley, R.J., Zhang, L., Lee, C.C., Schilkey, F.D., Sheth, V., Woodward, J.E., Peckham, H.E., Schroth, G.P., Kim, R.W., Kingsmore, S.F., 2011. Carrier testing for severe childhood recessive diseases by next-generation sequencing. Sci. Transl. Med. 3, 65ra4. doi:10.1126/scitranslmed.3001756

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R., others, 2008. Accurate whole human genome sequencing using reversible terminator chemistry. nature 456, 53–59.

Beulens, J.W., Grobbee, D.E., Nealb, B., others, 2010. The global burden of diabetes and its complications: an emerging pandemic. Eur. J. Cardiovasc. Prev. Rehabil. 17, s3–s8.

Biesecker, L.G., Green, R.C., 2014. Diagnostic clinical genome and exome sequencing. N. Engl. J. Med. 370, 2418–2425.

Bihl, G.R., Petri, M., Fine, D.M., 2006. Kidney biopsy in lupus nephritis: look before you leap. Nephrol. Dial. Transplant. 21, 1749–1752.

Bilgüvar, K., Oztürk, A.K., Louvi, A., Kwan, K.Y., Choi, M., Tatli, B., Yalnizoğlu, D., Tüysüz, B., Cağlayan, A.O., Gökben, S., Kaymakçalan, H., Barak, T., Bakircioğlu, M., Yasuno, K., Ho, W., Sanders, S., Zhu, Y., Yilmaz, S., Dinçer, A., Johnson, M.H., Bronen, R.A., Koçer, N., Per, H., Mane, S., Pamir, M.N., Yalçinkaya, C., Kumandaş, S., Topçu, M., Ozmen, M., Sestan, N., Lifton, R.P., State, M.W., Günel, M., 2010. Whole-exome sequencing identifies recessive WDR62 mutations in severe brain malformations. Nature 467, 207–210. doi:10.1038/nature09327

Bioinformatics, B., 2011. FASTQC: A quality control tool for high throughput sequence data. Cambridge, UK: Babraham Institute.

Bodmer, W., Bonilla, C., 2008. Common and rare variants in multifactorial susceptibility to common diseases. Nat. Genet. 40, 695–701. doi:10.1038/ng.f.136

Borch-Johnsen, K., Norgaard, K., Hommel, E., Mathiesen, E.R., Jensen, J.S., Deckert, T., Parving, H.-H., 1992. Is diabetic nephropathy an inherited complication. Kidney Int 41, 719–722.

Botstein, D., Risch, N., 2003. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nat. Genet. 33 Suppl, 228–237. doi:10.1038/ng1090

Botstein, D., White, R.L., Skolnick, M., Davis, R.W., 1980. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am. J. Hum. Genet. 32, 314.

Bowden, D.W., 2003. Genetics of kidney disease. Kidney Int. 63, S8–S12.

Boycott, K.M., Vanstone, M.R., Bulman, D.E., MacKenzie, A.E., 2013. Rare-disease genetics in the era of next-generation sequencing: discovery to translation. Nat. Rev. Genet. 14, 681–691.

Boyer, O., Benoit, G., Gribouval, O., Nevo, F., Tête, M.-J., Dantal, J., Gilbert-Dussardier, B., Touchard, G., Karras, A., Presne, C., others, 2011. Mutations in INF2 are a major cause of autosomal dominant focal segmental glomerulosclerosis. J. Am. Soc. Nephrol. 22, 239–245.

Brooke, E.M., others, 1974. The current and future use of registers in health information systems.

Brunham, L.R., Hayden, M.R., 2013. Hunting human disease genes: lessons from the past, challenges for the future. Hum. Genet. 132, 603–617.

Campbell, M.C., Tishkoff, S.A., 2008. AFRICAN GENETIC DIVERSITY: Implications for Human Demographic History, Modern Human Origins, and Complex Disease Mapping. Annu. Rev. Genomics Hum. Genet. 9, 403–433. doi:10.1146/annurev.genom.9.081307.164258

Carneiro, M.O., Russ, C., Ross, M.G., Gabriel, S.B., Nusbaum, C., DePristo, M.A., 2012. Pacific biosciences sequencing technology for genotyping and variation discovery in human data. BMC Genomics 13, 375.

Chahrour, M.H., Timothy, W.Y., Lim, E.T., Ataman, B., Coulter, M.E., Hill, R.S., Stevens, C.R., Schubert, C.R., Greenberg, M.E., Gabriel, S.B., others, 2012. Whole-exome sequencing and homozygosity analysis implicate depolarization-regulated neuronal genes in autism. PLoS Genet. 8, e1002635.

Chatterjee, R., Hoffman, M., Cliften, P., Seshan, S., Liapis, H., Jain, S., 2013. Targeted exome sequencing integrated with clinicopathological information reveals novel and rare mutations in atypical, suspected and unknown cases of Alport syndrome or proteinuria. PloS One 8, e76360.

Choi, M., Scholl, U.I., Ji, W., Liu, T., Tikhonova, I.R., Zumbo, P., Nayir, A., Bakkaloğlu, A., Özen, S., Sanjad, S., others, 2009a. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proc. Natl. Acad. Sci. 106, 19096–19101.

Choi, M., Scholl, U.I., Ji, W., Liu, T., Tikhonova, I.R., Zumbo, P., Nayir, A., Bakkalo\uglu, A., Özen, S., Sanjad, S., 2009b. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proc. Natl. Acad. Sci. 106, 19096–19101.

Chun, S., Fay, J.C., 2009. Identification of deleterious mutations within three human genomes. Genome Res. 19, 1553–1561.

Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., Ruden, D.M., 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly (Austin) 6, 80–92. doi:10.4161/fly.19695

Clark, M.J., Chen, R., Lam, H.Y., Karczewski, K.J., Chen, R., Euskirchen, G., Butte, A.J., Snyder, M., 2011. Performance comparison of exome DNA sequencing technologies. Nat. Biotechnol. 29, 908–914.

Consortium, 1000 Genomes Project, others, 2012. An integrated map of genetic variation from 1,092 human genomes. Nature 491, 56–65.

Consortium, 1000 Genomes Project, others, 2011. A map of human genome variation

from population-scale sequencing. Nature 473, 544–544.
Consortium, 1000 Genomes Project, others, 2010. A map of human genome variation from population-scale sequencing. Nature 467, 1061–1073.
Cooper, G.M., Shendure, J., 2011. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nat. Rev. Genet. 12, 628–640. doi:10.1038/nrg3046
Correa-Rotter, R., Gonzalez-Michaca, L., 2005. Early detection and prevention of diabetic nephropathy: a challenge calling for mandatory action for Mexico and the developing world. Kidney Int. 68, S69–S75.
Covic, A., Schiller, A., Volovat, C., Gluhovschi, G., Gusbeth-Tatomir, P., Petrica, L., Caruntu, I.-D., Bozdog, G., Velciov, S., Trandafirescu, V., others, 2006. Epidemiology of renal disease in Romania: a 10 year review of two regional renal biopsy databases. Nephrol. Dial. Transplant. 21, 419–424.
Danchenko, N., Satia, J.A., Anthony, M.S., 2006. Epidemiology of systemic lupus erythematosus: a comparison of worldwide disease burden. Lupus 15, 308–318.
Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., others, 2011. The variant call format and VCFtools. Bioinformatics 27, 2156–2158.
Davydov, E.V., Goode, D.L., Sirota, M., Cooper, G.M., Sidow, A., Batzoglou, S., 2010. Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput. Biol. 6, e1001025.
Day-Williams, A.G., Zeggini, E., 2011. The effect of next-generation sequencing technology on complex trait research. Eur. J. Clin. Invest. 41, 561–567.
De Baets, G., Van Durme, J., Reumers, J., Maurer-Stroh, S., Vanhee, P., Dopazo, J., Schymkowitz, J., Rousseau, F., 2011. SNPeffect 4.0: on-line prediction of molecular and structural effects of protein-coding variants. Nucleic Acids Res. gkr996.
Deltas, C., Pierides, A., Voskarides, K., 2013. Molecular genetics of familial hematuric diseases. Nephrol. Dial. Transplant. 28, 2946–2960.
DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., others, 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 43, 491–498.
Dhaun, N., Bellamy, C.O., Cattran, D.C., Kluth, D.C., 2014. Utility of renal biopsy in the clinical management of renal disease. Kidney Int.
Dixon-Salazar, T.J., Silhavy, J.L., Udpa, N., Schroth, J., Bielas, S., Schaffer, A.E., Olvera, J., Bafna, V., Zaki, M.S., Abdel-Salam, G.H., others, 2012. Exome sequencing can improve diagnosis and alter patient management. Sci. Transl. Med. 4, 138ra78–138ra78.
Dreyer, S.D., Zhou, G., Baldini, A., Winterpacht, A., Zabel, B., Cole, W., Johnson, R.L., Lee, B., 1998. Mutations in LMX1B cause abnormal skeletal patterning and renal dysplasia in nail patella syndrome. Nat. Genet. 19, 47–50.
DuBose, T.D., 2007. American Society of Nephrology Presidential Address 2006: chronic kidney disease as a public health threat—new strategy for a growing problem. J. Am. Soc. Nephrol. 18, 1038–1045.
Edwards, N., Rice, S.J., Raman, S., Hynes, A.M., Srivastava, S., Moore, I., Al-Hamed, M., Xu, Y., Santibanez-Koref, M., Thwaites, D.T., others, 2014. A novel LMX1B mutation in a family with end-stage renal disease of “unknown cause.” Clin. Kidney J. sfu129.

Ellsworth, R., Howard, J.E., 1934. Studies on the physiology of the parathyroid glands. VII. Some responses of normal human kidneys and blood to intravenous parathyroid extract. Bull Johns Hopkins Hosp 55, 195.

Fallerini, C., Dosa, L., Tita, R., Del Prete, D., Feriozzi, S., Gai, G., Clementi, M., La Manna, A., Miglietti, N., Mancini, R., others, 2014. Unbiased next generation sequencing analysis confirms the existence of autosomal dominant Alport syndrome in a relevant fraction of cases. Clin. Genet. 86, 252–257.

Feder, J.N., Gnirke, A., Thomas, W., Tsuchihashi, Z., Ruddy, D.A., Basava, A., Dormishian, F., Domingo, R., Ellis, M.C., Fullan, A., others, 1996. A novel MHC class I–like gene is mutated in patients with hereditary haemochromatosis. Nat. Genet. 13, 399–408.

Frasca, G.M., Soverini, L., Gharavi, A.G., Lifton, R.P., Canova, C., Preda, P., Vangelista, A., Stefoni, S., 2004. Thin basement membrane disease in patients with familial IgA nephropathy. J Nephrol 17, 778–785.

Freedman, B.I., Kopp, J.B., Langefeld, C.D., Genovese, G., Friedman, D.J., Nelson, G.W., Winkler, C.A., Bowden, D.W., Pollak, M.R., 2010. The apolipoprotein L1 (APOL1) gene and nondiabetic nephropathy in African Americans. J. Am. Soc. Nephrol. 21, 1422–1426.

Freedman, B.I., Spray, B.J., Tuttle, A.B., Buckalew Jr, V.M., 1993. The familial risk of end-stage renal disease in African Americans. Am. J. Kidney Dis. 21, 387–393.

Furth, S.L., Cole, S.R., Moxey-Mims, M., Kaskel, F., Mak, R., Schwartz, G., Wong, C., Muñoz, A., Warady, B.A., 2006. Design and methods of the Chronic Kidney Disease in Children (CKiD) prospective cohort study. Clin. J. Am. Soc. Nephrol. 1, 1006–1015.

Gayà-Vidal, M., Albà, M.M., 2014. Uncovering adaptive evolution in the human lineage. BMC Genomics 15, 599. doi:10.1186/1471-2164-15-599

Ge, D., Ruzzo, E.K., Shianna, K.V., He, M., Pelak, K., Heinzen, E.L., Need, A.C., Cirulli, E.T., Maia, J.M., Dickson, S.P., others, 2011. SVA: software for annotating and visualizing sequenced human genomes. Bioinformatics 27, 1998–2000.

Gee, H.Y., Otto, E.A., Hurd, T.W., Ashraf, S., Chaki, M., Cluckey, A., Vega-Warner, V., Saisawat, P., Diaz, K.A., Fang, H., others, 2013. Whole-exome resequencing distinguishes cystic kidney diseases from phenocopies in renal ciliopathies. Kidney Int. 85, 880–887.

Genovese, F., Manresa, A.A., Leeming, D.J., Karsdal, M.A., Boor, P., 2014. The extracellular matrix in the kidney: a source of novel non-invasive biomarkers of kidney fibrosis? Fibrogenesis Tissue Repair 7.

Genovese, G., Friedman, D.J., Ross, M.D., Lecordier, L., Uzureau, P., Freedman, B.I., Bowden, D.W., Langefeld, C.D., Oleksyk, T.K., Knob, A.L.U., others, 2010. Association of trypanolytic ApoL1 variants with kidney disease in African Americans. Science 329, 841–845.

George Jr, A.L., Neilson, E.G., 2000. Genetics of kidney disease. Am. J. Kidney Dis. 35, S160–S169.

Ghoul, B. El, Squalli, T., Servais, A., Elie, C., Meas-Yedid, V., Trivint, C., Vanmassenhove, J., Grünfeld, J.-P., Olivo-Marin, J.-C., Thervet, E., others, 2010. Urinary procollagen III aminoterminal propeptide (PIIINP): a fibrotest for the nephrologist. Clin. J. Am. Soc. Nephrol. 5, 205–210.

Gibson, J., Gilbert, R.D., BUNYAN, D.J., Angus, E.M., Fowler, D.J., Ennis, S., 2013. Exome analysis resolves differential diagnosis of familial kidney disease and

uncovers a potential confounding variant. Genet. Res. 95, 165–173.
Gipson, D.S., Chin, H., Presler, T.P., Jennette, C., Ferris, M.E., Massengill, S., Gibson, K., Thomas, D.B., 2006. Differential risk of remission and ESRD in childhood FSGS. Pediatr. Nephrol. 21, 344–349.
Girard, S.L., Gauthier, J., Noreau, A., Xiong, L., Zhou, S., Jouan, L., Dionne-Laporte, A., Spiegelman, D., Henrion, E., Diallo, O., Thibodeau, P., Bachand, I., Bao, J.Y.J., Tong, A.H.Y., Lin, C.-H., Millet, B., Jaafari, N., Joober, R., Dion, P.A., Lok, S., Krebs, M.-O., Rouleau, G.A., 2011. Increased exonic de novo mutation rate in individuals with schizophrenia. Nat. Genet. 43, 860–863. doi:10.1038/ng.886
Gladman, D.D., UROWITZ, M.B., COLE, E., RITCHIE, S., CHANG, C.H., CHURG, J., 1989. Kidney biopsy in SLE. I. A clinical-morphologic evaluation. QJM 73, 1125–1133.
Gliklich, R., Dreyer, N., 2010. Registries for Evaluating Patient Outcomes. Agency Healthc. Res. Qual. Rockv. MD.
Goldstein, J.L., Hobbs, H.H., Brown, M.S., 2001. The metabolic and molecular bases of inherited disease. Fam. Hypercholesterolemia N. Y. McGraw-Hill 2863–913.
Gomez, F., Hirbo, J., Tishkoff, S.A., 2014. Genetic Variation and Adaptation in Africa: Implications for Human Evolution and Disease. Cold Spring Harb. Perspect. Biol. 6, a008524. doi:10.1101/cshperspect.a008524
Gonzaga-Jauregui, C., Lupski, J.R., Gibbs, R.A., 2012. Human genome sequencing in health and disease. Annu. Rev. Med. 63, 35.
Gonzalez-Angulo, A.M., Hennessy, B.T., Mills, G.B., 2010. Future of personalized medicine in oncology: a systems biology approach. J. Clin. Oncol. 28, 2777–2783.
Grada, A., Weinbrecht, K., 2013. Next-Generation Sequencing: Methodology and Application. J. Invest. Dermatol. 133, e11. doi:10.1038/jid.2013.248
Green, R.C., Rehm, H.L., Kohane, I.S., 2013. Clinical genome sequencing. Genomic Pers. Med. 2.
Gross, O., Netzer, K.-O., Lambrecht, R., Seibold, S., Weber, M., 2002. Meta-analysis of genotype–phenotype correlation in X-linked Alport syndrome: impact on clinical counselling. Nephrol. Dial. Transplant. 17, 1218–1227.
Gymrek, M., Golan, D., Rosset, S., Erlich, Y., 2012. lobSTR: A short tandem repeat profiler for personal genomes. Genome Res. 22, 1154–1162. doi:10.1101/gr.135780.111
Haider, D.G., Friedl, A., Peric, S., Wiesinger, G.F., Wolzt, M., Prosenz, J., Fischer, H., Hörl, W.H., Soleiman, A., Fuhrmann, V., 2012. Kidney biopsy in patients with glomerulonephritis: is the earlier the better? BMC Nephrol. 13, 34.
Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A., McKusick, V.A., 2005. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res. 33, D514–D517.
Hansen, J., Snow, C., Tuttle, E., Ghoneim, D.H., Yang, C.-S., Spencer, A., Gunter, S.A., Smyser, C.D., Gurnett, C.A., Shinawi, M., Dobyns, W.B., Wheless, J., Halterman, M.W., Jansen, L.A., Paschal, B.M., Paciorkowski, A.R., 2015. De Novo Mutations in SIK1 Cause a Spectrum of Developmental Epilepsies. Am. J. Hum. Genet. 96, 682–690. doi:10.1016/j.ajhg.2015.02.013
Harris, P.A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., Conde, J.G., 2009. Research electronic data capture (REDCap)—A metadata-driven methodology and workflow process for providing translational research informatics support.

J. Biomed. Inform. 42, 377–381. doi:10.1016/j.jbi.2008.08.010
Herzog, C.A., Ma, J.Z., Collins, A.J., 2002. Long-Term Survival of Dialysis Patients in the United States With Prosthetic Heart Valves Should ACC/AHA Practice Guidelines on Valve Selection Be Modified? Circulation 105, 1336–1341.
Hildebrandt, F., Strahm, B., Nothwang, H.-G., Gretz, N., Schnieders, B., Singh-Sawhney, I., Kutt, R., Vollmer, M., Brandis, M., 1997. Molecular genetic identification of families with juvenile nephronophthisis type 1: rate of progression to renal failure. Kidney Int. 51, 261–269.
Hillyer, M., 2010. Managing Hierarchical Data in MySQL. MySQL Dev. Zone Online.
Hodkinson, B., Mapiye, D., Jayne, D., Kalla, A., Tiffin, N., Okpechi, I., 2015. The African Lupus Genetics Network (ALUGEN) registry: standardized, prospective follow-up studies in African patients with systemic lupus erythematosus. Lupus 0961203315606984.
Hossain, M.P., Goyder, E.C., Rigby, J.E., Nahas, M. El, 2009. CKD and poverty: a growing global challenge. Am. J. Kidney Dis. 53, 166–174.
Hudson, B.G., 2004. The molecular basis of Goodpasture and Alport syndromes: beacons for the discovery of the collagen IV family. J. Am. Soc. Nephrol. 15, 2514–2527.
Hunt, K.A., Smyth, D.J., Balschun, T., Ban, M., Mistry, V., Ahmad, T., Anand, V., Barrett, J.C., Bhaw-Rosun, L., Bockett, N.A., Brand, O.J., Brouwer, E., Concannon, P., Cooper, J.D., Dias, K.-R.M., van Diemen, C.C., Dubois, P.C., Edkins, S., Fölster-Holst, R., Fransen, K., Glass, D.N., Heap, G.A.R., Hofmann, S., Huizinga, T.W.J., Hunt, S., Langford, C., Lee, J., Mansfield, J., Marrosu, M.G., Mathew, C.G., Mein, C.A., Müller-Quernheim, J., Nutland, S., Onengut-Gumuscu, S., Ouwehand, W., Pearce, K., Prescott, N.J., Posthumus, M.D., Potter, S., Rosati, G., Sambrook, J., Satsangi, J., Schreiber, S., Shtir, C., Simmonds, M.J., Sudman, M., Thompson, S.D., Toes, R., Trynka, G., Vyse, T.J., Walker, N.M., Weidinger, S., Zhernakova, A., Zoledziewska, M., Type 1 Diabetes Genetics Consortium, UK Inflammatory Bowel Disease (IBD) Genetics Consortium, Wellcome Trust Case Control Consortium, Weersma, R.K., Gough, S.C.L., Sawcer, S., Wijmenga, C., Parkes, M., Cucca, F., Franke, A., Deloukas, P., Rich, S.S., Todd, J.A., van Heel, D.A., 2012. Rare and functional SIAE variants are not associated with autoimmune disease risk in up to 66,924 individuals of European ancestry. Nat. Genet. 44, 3–5. doi:10.1038/ng.1037
Imperatore, G., Hanson, R.L., Pettitt, D.J., Kobes, S., Bennett, P.H., Knowler, W.C., 1998. Sib-pair linkage analysis for susceptibility genes for microvascular complications among Pima Indians with type 2 diabetes. Pima Diabetes Genes Group. Diabetes 47, 821–830.
Ingram, G.I., 1976. The history of haemophilia. J. Clin. Pathol. 29, 469–479.
Iyengar, S.K., Abboud, H.E., Goddard, K.A., Saad, M.F., Adler, S.G., Arar, N.H., Bowden, D.W., Duggirala, R., Elston, R.C., Hanson, R.L., others, 2007. Genome-Wide Scans for Diabetic Nephropathy and Albuminuria in Multiethnic Populations The Family Investigation of Nephropathy and Diabetes (FIND). Diabetes 56, 1577–1585.
Jaar, B.G., Coresh, J., Plantinga, L.C., Fink, N.E., Klag, M.J., Levey, A.S., Levin, N.W., Sadler, J.H., Kliger, A., Powe, N.R., 2005. Comparing the risk for death with peritoneal dialysis and hemodialysis in a national cohort of patients with chronic kidney disease. Ann. Intern. Med. 143, 174–183.

Jafar, T.H., Islam, M., Poulter, N., others, 2006. Chronic kidney disease in the developing world.

Jha, V., Garcia-Garcia, G., Iseki, K., Li, Z., Naicker, S., Plattner, B., Saran, R., Wang, A.Y.-M., Yang, C.-W., 2013. Chronic kidney disease: global dimension and perspectives. The Lancet 382, 260–272.

Kao, W.L., Klag, M.J., Meoni, L.A., Reich, D., Berthier-Schaad, Y., Li, M., Coresh, J., Patterson, N., Tandon, A., Powe, N.R., others, 2008. MYH9 is associated with nondiabetic end-stage renal disease in African Americans. Nat. Genet. 40, 1185–1192.

Katz, I.J., Gerntholtz, T., Naicker, S., 2010. Africa and nephrology: the forgotten continent. Nephron Clin. Pract. 117, 320–327.

Kelly, B.J., Fitch, J.R., Hu, Y., Corsmeier, D.J., Zhong, H., Wetzel, A.N., Nordquist, R.D., Newsom, D.L., White, P., 2015. Churchill: an ultra-fast, deterministic, highly scalable and balanced parallelization strategy for the discovery of human genetic variation in clinical and population-scale genomics. Genome Biol. 16, 6.

Kerem, B., Rommens, J.M., Buchanan, J.A., Markiewicz, D., Cox, T.K., Chakravarti, A., Buchwald, M., Tsui, L.-C., 1989. Identification of the cystic fibrosis gene: genetic analysis. Science 245, 1073–1080.

Khazen, D., Jendoubi-Ayed, S., Gorgi, Y., Sfar, I., Abderrahim, E., Abdallah, T.B., Ayed, K., 2007. Adhesion molecule polymorphisms in acute renal allograft rejection, in: Transplantation Proceedings. Elsevier, pp. 2563–2564.

Kirby, A., Gnirke, A., Jaffe, D.B., Barešová, V., Pochet, N., Blumenstiel, B., Ye, C., Aird, D., Stevens, C., Robinson, J.T., others, 2013. Mutations causing medullary cystic kidney disease type 1 lie in a large VNTR in MUC1 missed by massively parallel sequencing. Nat. Genet. 45, 299–303.

Koboldt, D.C., Zhang, Q., Larson, D.E., Shen, D., McLellan, M.D., Lin, L., Miller, C.A., Mardis, E.R., Ding, L., Wilson, R.K., 2012. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 22, 568–576.

Kopple, J.D., 2001. National kidney foundation K/DOQI clinical practice guidelines for nutrition in chronic renal failure. Am. J. Kidney Dis. 37, S66–S70.

Kottgen, A., Kao, W.H., Hwang, S.-J., Boerwinkle, E., Yang, Q., Levy, D., Benjamin, E.J., Larson, M.G., Astor, B.C., Coresh, J., others, 2008. Genome-wide association study for renal traits in the Framingham Heart and Atherosclerosis Risk in Communities Studies. BMC Med. Genet. 9, 49.

Krumm, N., Sudmant, P.H., Ko, A., O'Roak, B.J., Malig, M., Coe, B.P., NHLBI Exome Sequencing Project, Quinlan, A.R., Nickerson, D.A., Eichler, E.E., 2012. Copy number variation detection and genotyping from exome sequence data. Genome Res. 22, 1525–1532. doi:10.1101/gr.138115.112

Kumar, A., White, T.A., MacKenzie, A.P., Clegg, N., Lee, C., Dumpit, R.F., Coleman, I., Ng, S.B., Salipante, S.J., Rieder, M.J., others, 2011. Exome sequencing identifies a spectrum of mutation frequencies in advanced and lethal prostate cancers. Proc. Natl. Acad. Sci. 108, 17087–17092.

Kumar, N., 2013. SAS Tricks: Recently updated Base SAS Certification question and answer. SAS Tricks.

Kuo, D.S., Labelle-Dumais, C., Gould, D.B., 2012. COL4A1 and COL4A2 mutations and disease: insights into pathogenic mechanisms and potential therapeutic targets. Hum. Mol. Genet. 21, R97–R110. doi:10.1093/hmg/dds346

Lander, E.S., 2011. Initial impact of the sequencing of the human genome. Nature

470, 187–197.
Lander, E.S., Botstein, D., 1987. Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. Science 236, 1567–1570.
Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359.
Lee, H., Deignan, J.L., Dorrani, N., Strom, S.P., Kantarci, S., Quintero-Rivera, F., Das, K., Toy, T., Harry, B., Yourshaw, M., others, 2014. Clinical exome sequencing for genetic identification of rare mendelian disorders. JAMA 312, 1880–1887.
Liew, W.K., Ben-Omran, T., Darras, B.T., Prabhu, S.P., Darryl, C., Vatta, M., Yang, Y., Eng, C.M., Chung, W.K., 2013. Clinical application of whole-exome sequencing: a novel autosomal recessive spastic ataxia of Charlevoix-Saguenay sequence variation in a child with ataxia. JAMA Neurol. 70, 788–791.
Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25, 1754–1760.
Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., others, 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25, 2078–2079.
Li, H., Homer, N., 2010. A survey of sequence alignment algorithms for next-generation sequencing. Brief. Bioinform. 11, 473–483.
LIM, T.-O., Goh, A., LIM, Y.-N., Morad, Z., 2008. Review article: Use of renal registry data for research, health-care planning and quality improvement: What can we learn from registry data in the Asia–Pacific region? Nephrology 13, 745–752.
Lin, F., Bian, F., Zou, J., Wu, X., Shan, J., Lu, W., Yao, Y., Jiang, G., Gale, D.P., 2014. Whole exome sequencing reveals novel COL4A3 and COL4A4 mutations and resolves diagnosis in Chinese families with kidney disease. BMC Nephrol. 15, 175.
Li, R., Li, Y., Fang, X., Yang, H., Wang, J., Kristiansen, K., Wang, J., 2009. SNP detection for massively parallel whole-genome resequencing. Genome Res. 19, 1124–1132.
Liu, D.J., Leal, S.M., 2010. A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions. PLoS Genet. 6, e1001156.
Loraine, A.E., Helt, G.A., 2002. Visualizing the genome: techniques for presenting human genome data and annotations. BMC Bioinformatics 3, 19.
Lu, L.-J., Wallace, D.J., Navarra, S.V., Weisman, M.H., 2010. Lupus registries: evolution and challenges, in: Seminars in Arthritis and Rheumatism. Elsevier, pp. 224–245.
Lupski, J.R., Gonzaga-Jauregui, C., Yang, Y., Bainbridge, M.N., Jhangiani, S., Buhay, C.J., Kovar, C.L., Wang, M., Hawes, A.C., Reid, J.G., others, 2013. Exome sequencing resolves apparent incidental findings and reveals further complexity of SH3TC2 variant alleles causing Charcot-Marie-Tooth neuropathy. Genome Med 5, 57.
Mabayoje, M.O., Bamgboye, E.L., Odutola, T.A., Mabadeje, A.F.B., 1992. Chronic renal failure at the Lagos University Teaching Hospital: a 10-year review, in: Transplantation Proceedings. Elsevier, pp. 1851–1852.
MacArthur, D.G., Balasubramanian, S., Frankish, A., Huang, N., Morris, J., Walter, K., Jostins, L., Habegger, L., Pickrell, J.K., Montgomery, S.B., Albers, C.A.,

Zhang, Z.D., Conrad, D.F., Lunter, G., Zheng, H., Ayub, Q., DePristo, M.A., Banks, E., Hu, M., Handsaker, R.E., Rosenfeld, J.A., Fromer, M., Jin, M., Mu, X.J., Khurana, E., Ye, K., Kay, M., Saunders, G.I., Suner, M.-M., Hunt, T., Barnes, I.H.A., Amid, C., Carvalho-Silva, D.R., Bignell, A.H., Snow, C., Yngvadottir, B., Bumpstead, S., Cooper, D.N., Xue, Y., Romero, I.G., 1000 Genomes Project Consortium, Wang, J., Li, Y., Gibbs, R.A., McCarroll, S.A., Dermitzakis, E.T., Pritchard, J.K., Barrett, J.C., Harrow, J., Hurles, M.E., Gerstein, M.B., Tyler-Smith, C., 2012. A systematic survey of loss-of-function variants in human protein-coding genes. Science 335, 823–828. doi:10.1126/science.1215040

Magi, A., Benelli, M., Gozzini, A., Girolami, F., Torricelli, F., Brandi, M.L., 2010. Bioinformatics for next generation sequencing data. Genes 1, 294–307.

Maisonneuve, P., Agodoa, L., Gellert, R., Stewart, J.H., Buccianti, G., Lowenfels, A.B., Wolfe, R.A., Jones, E., Disney, A.P., Briggs, D., others, 2000. Distribution of primary renal diseases leading to end-stage renal failure in the United States, Europe, and Australia/New Zealand: results from an international comparative study. Am. J. Kidney Dis. 35, 157–165.

Majewski, J., Wang, Z., Lopez, I., Humaid, S. Al, Ren, H., Racine, J., Bazinet, A., Mitchel, G., Braverman, N., Koenekoop, R.K., 2011. A new ocular phenotype associated with an unexpected but known systemic disorder and mutation: novel use of genomic diagnostics and exome sequencing. J. Med. Genet. 48, 593–596.

Malone, A.F., Phelan, P.J., Hall, G., Cetincelik, U., Homstad, A., Alonso, A.S., Jiang, R., Lindsey, T.B., Wu, G., Sparks, M.A., others, 2014. Rare hereditary COL4A3/COL4A4 variants may be mistaken for familial focal segmental glomerulosclerosis. Kidney Int. 86, 1253–1259.

Mardis, E.R., 2011. A decade/'s perspective on DNA sequencing technology. Nature 470, 198–203.

Martins, D., Agodoa, L., Norris, K., 2012. Chronic kidney disease in disadvantaged populations. Int. J. Nephrol. 2012.

Matekole, M., Affram, K., Lee, S.J., Howie, A.J., Michael, J., Adu, D., 1993. Hypertension and end-stage renal failure in tropical Africa. J. Hum. Hypertens. 7, 443–446.

Mboowa, G., 2014. Genetics of Sub-Saharan African Human Population: Implications for HIV/AIDS, Tuberculosis, and Malaria. Int. J. Evol. Biol. 2014, e108291. doi:10.1155/2014/108291

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., DePristo, M.A., 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20, 1297–1303. doi:10.1101/gr.107524.110

Mclaren, A.J., Marshall, S.E., Haldar, N.A., Mullighan, C.G., Fuggle, S.V., Morris, P.J., Welsh, K.I., 1999. Adhesion molecule polymorphisms in chronic renal allograft failure. Kidney Int. 55, 1977–1982.

Medina, I., De Maria, A., Bleda, M., Salavert, F., Alonso, R., Gonzalez, C.Y., Dopazo, J., 2012. VARIANT: Command Line, Web service and Web interface for fast and accurate functional characterization of variants found by Next-Generation Sequencing. Nucleic Acids Res. 40, W54–W58.

Mensah, G.A., Mayosi, B.M., 2013. The 2011 United Nations high-level meeting on non-communicable diseases: the Africa agenda calls for a 5-by-5 approach.

SAMJ South Afr. Med. J. 103, 77–79.
Metzker, M.L., 2010. Sequencing technologies—the next generation. Nat. Rev. Genet. 11, 31–46.
Meynert, A.M., Ansari, M., FitzPatrick, D.R., Taylor, M.S., 2014. Variant detection sensitivity and biases in whole genome and exome sequencing. BMC Bioinformatics 15, 247.
Molho-Pessach, V., Rios, J.J., Xing, C., Setchell, K.D., Cohen, J.C., Hobbs, H.H., 2012. Homozygosity mapping identifies a bile acid biosynthetic defect in an adult with cirrhosis of unknown etiology. Hepatology 55, 1139–1145.
Morinière, V., Dahan, K., Hilbert, P., Lison, M., Lebbah, S., Topa, A., Bole-Feysot, C., Pruvost, S., Nitschke, P., Plaisier, E., others, 2014. Improving Mutation Screening in Familial Hematuric Nephropathies through Next Generation Sequencing. J. Am. Soc. Nephrol. ASN–2013080912.
Murray, C.J., Lopez, A.D., 1997. Alternative projections of mortality and disability by cause 1990–2020: Global Burden of Disease Study. The Lancet 349, 1498–1504.
MySQL, A.B., 1997. MySQL developer zone. Sun Microsyst. C1995-2008 Cit 2009-05-25 Kódováno 8.
Naicker, S., 2010. Burden of end-stage renal disease in sub-Saharan Africa. Clin. Nephrol. 74, S13–6.
Naicker, S., 2003. End-stage renal disease in sub-Saharan and South Africa. Kidney Int. 63, S119–S122.
NAICKER, S., 1998. Patterns of renal disease in South Africa. Nephrology 4, S21–S24.
Need, A.C., Shashi, V., Hitomi, Y., Schoch, K., Shianna, K.V., McDonald, M.T., Meisler, M.H., Goldstein, D.B., 2012. Clinical application of exome sequencing in undiagnosed genetic conditions. J. Med. Genet. jmedgenet–2012.
Network, C.G.A., others, 2012. Comprehensive molecular portraits of human breast tumours. Nature 490, 61–70.
Ng, P.C., Henikoff, S., 2003. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res. 31, 3812–3814.
Ng, P.C., Levy, S., Huang, J., Stockwell, T.B., Walenz, B.P., Li, K., Axelrod, N., Busam, D.A., Strausberg, R.L., Venter, J.C., 2008. Genetic Variation in an Individual Human Exome. PLoS Genet 4, e1000160. doi:10.1371/journal.pgen.1000160
Ng, S.B., Bigham, A.W., Buckingham, K.J., Hannibal, M.C., McMillin, M.J., Gildersleeve, H.I., Beck, A.E., Tabor, H.K., Cooper, G.M., Mefford, H.C., 2010. Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. Nat. Genet. 42, 790–793.
Ng, S.B., Turner, E.H., Robertson, P.D., Flygare, S.D., Bigham, A.W., Lee, C., Shaffer, T., Wong, M., Bhattacharjee, A., Eichler, E.E., Bamshad, M., Nickerson, D.A., Shendure, J., 2009a. Targeted capture and massively parallel sequencing of 12 human exomes. Nature 461, 272–276. doi:10.1038/nature08250
Ng, S.B., Turner, E.H., Robertson, P.D., Flygare, S.D., Bigham, A.W., Lee, C., Shaffer, T., Wong, M., Bhattacharjee, A., Eichler, E.E., others, 2009b. Targeted capture and massively parallel sequencing of 12 human exomes. Nature 461, 272–276.
Nishisho, I., Nakamura, Y., Miyoshi, Y., Miki, Y., Ando, H., Horii, A., Koyama, K.,

Utsunomiya, J., Baba, S., Hedge, P., 1991. Mutations of chromosome 5q21 genes in FAP and colorectal cancer patients. Science 253, 665–669.

Norton, N., Robertson, P.D., Rieder, M.J., Züchner, S., Rampersaud, E., Martin, E., Li, D., Nickerson, D.A., Hershberger, R.E., National Heart, Lung and Blood Institute GO Exome Sequencing Project, 2012. Evaluating pathogenicity of rare variants from dilated cardiomyopathy in the exome era. Circ. Cardiovasc. Genet. 5, 167–174. doi:10.1161/CIRCGENETICS.111.961805

Nossent, I.C., Henzen-Logmans, S.C., Vroom, T.M., Huysen, V., Berden, J.H.M., Swaak, A.J.G., 1991. Relation between serological data at the time of biopsy and renal histology in lupus nephritis. Rheumatol. Int. 11, 77–82.

Nugent, R.A., Fathima, S.F., Feigl, A.B., Chyung, D., 2011. The burden of chronic kidney disease on developing nations: a 21st century challenge in global health. Nephron Clin. Pract. 118, c269–c277.

O'Dea, D.F., Murphy, S.W., Hefferton, D., Parfrey, P.S., 1998. Higher risk for renal failure in first-degree relatives of white patients with end-stage renal disease: a population-based study. Am. J. Kidney Dis. 32, 794–801.

Odubanjo, M.O., Oluwasola, A.O., Kadiri, S., 2011. The epidemiology of end-stage renal disease in Nigeria: the way forward. Int. Urol. Nephrol. 43, 785–792.

Okada, Y., Sim, X., Go, M.J., Wu, J.-Y., Gu, D., Takeuchi, F., Takahashi, A., Maeda, S., Tsunoda, T., Chen, P., others, 2012. Meta-analysis identifies multiple loci associated with kidney function-related traits in east Asian populations. Nat. Genet. 44, 904–909.

Okpechi, I., Swanepoel, C., Duffield, M., Mahala, B., Wearne, N., Alagbe, S., Barday, Z., Arendse, C., Rayner, B., 2010. Patterns of renal disease in Cape Town South Africa: a 10-year review of a single-centre renal biopsy database. Nephrol. Dial. Transplant. gfq655.

Ong, A.C., Fine, L.G., 1994. Tubular-derived growth factors and cytokines in the pathogenesis of tubulointerstitial fibrosis: implications for human renal disease progression. Am. J. Kidney Dis. 23, 205–209.

Organization, W.H., others, 2005. Preventing chronic diseases: a vital investment.

Overduin, B., 2011. Determining the Effect of Genomic Variants Using the Ensembl Variant Effect Predictor [WWW Document]. Bioinforma. Knowledgeblog. URL http://bioinformatics.knowledgeblog.org/2011/06/21/determining-the-effect-of-genomic-variants-using-the-ensembl-variant-effect-predictor/ (accessed 8.12.15).

Pakistani, A., 1994. The growing burden of chronic kidney disease in Pakistan. Cornell Law Rev 79, 1382–404.

Pangrazio, A., Puddu, A., Oppo, M., Valentini, M., Zammataro, L., Vellodi, A., Gener, B., Llano-Rivas, I., Raza, J., Atta, I., others, 2014. Exome sequencing identifies< i> CTSK</i> mutations in patients originally diagnosed as intermediate osteopetrosis. Bone 59, 122–126.

Park, G., Gim, J., Kim, A.R., Han, K.-H., Kim, H.-S., Oh, S.-H., Park, T., Park, W.-Y., Choi, B.Y., 2013. Multiphasic analysis of whole exome sequencing data identifies a novel mutation of ACTG1 in a nonsyndromic hearing loss family. BMC Genomics 14, 191.

Pettersson, E., Lundeberg, J., Ahmadian, A., 2009. Generations of sequencing technologies. Genomics 93, 105–111.

Pettitt, D.J., Saad, M.F., Bennett, P.H., Nelson, R.G., Knowler, W.C., 1990. Familial predisposition to renal disease in two generations of Pima Indians with type 2 (non-insulin-dependent) diabetes mellitus. Diabetologia 33, 438–443.

Pierides, A., Voskarides, K., Athanasiou, Y., Ioannou, K., Damianou, L., Arsali, M., Zavros, M., Pierides, M., Vargemezis, V., Patsias, C., others, 2009. Clinico-pathological correlations in 127 patients in 11 large pedigrees, segregating one of three heterozygous mutations in the COL4A3/COL4A4 genes associated with familial haematuria and significant late progression to proteinuria and chronic kidney disease from focal segmental glomerulosclerosis. Nephrol. Dial. Transplant. gfp158.

Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R., Siepel, A., 2010. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 20, 110–121.

Porta, M.S., Greenland, S., Hernán, M., Silva, I.D.S., Last, J.M., 2014. A dictionary of epidemiology. Oxford University Press.

Pugsley, D., Norris, K.C., Garcia-Garcia, G., Agodoa, L., 2009. GLOBAL APPROACHES FOR UNDERSTANDING THE DISPROPORTIONATE BURDEN OF CHRONIC KIDNEY DISEASE. Ethn. Dis. 19, 1–S1.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., Sham, P.C., 2007. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am. J. Hum. Genet. 81, 559–575.

Ranganath, P., Tripathi, G., Sharma, R.K., Sankhwar, S.N., Agrawal, S., 2009. Role of non-HLA genetic variants in end-stage renal disease. Tissue Antigens 74, 147–155.

Rao, P.S., Merion, R.M., Ashby, V.B., Port, F.K., Wolfe, R.A., Kayler, L.K., 2007. Renal transplantation in elderly patients older than 70 years of age: results from the Scientific Registry of Transplant Recipients. Transplantation 83, 1069–1074.

Reyes-Bahamonde, J., Raimann, J.G., Canaud, B., Etter, M., Kooman, J.P., Levin, N.W., Marcelli, D., Marelli, C., Power, A., Van Der Sande, F.M., others, 2014. CKD GENERAL AND CLINICAL EPIDEMIOLOGY 1. Nephrol. Dial. Transplant. 29, iii124–iii139.

Robinson, P.N., Krawitz, P., Mundlos, S., 2011. Strategies for exome and genome sequence data analysis in disease-gene discovery projects. Clin. Genet. 80, 127–132.

Rozario, T., DeSimone, D.W., 2010. The extracellular matrix in development and morphogenesis: a dynamic view. Dev. Biol. 341, 126–140.

Ruffalo, M., LaFramboise, T., Koyutürk, M., 2011. Comparative analysis of algorithms for next-generation sequencing read alignment. Bioinformatics 27, 2790–2796.

Sadee, W., Hartmann, K., Seweryn, M., Pietrzak, M., Handelman, S.K., Rempala, G.A., 2014. Missing heritability of common diseases and treatments outside the protein-coding exome. Hum. Genet. 133, 1199–1215. doi:10.1007/s00439-014-1476-7

Saghir, N.S. El, Khalil, M.K., Eid, T., Kinge, A.R. El, Charafeddine, M., Geara, F., Seoud, M., Shamseddine, A.I., 2007. Trends in epidemiology and management of breast cancer in developing Arab countries: a literature and registry analysis. Int. J. Surg. 5, 225–233.

Sam Lim, S., Drenkard, C., McCune, W.J., Helmick, C.G., Gordon, C., DeGuire, P., Bayakly, R., Somers, E.C., 2009. Population-based lupus registries: Advancing our epidemiologic understanding. Arthritis Care Res. 61, 1462–1466.

Sankaran, V.G., Ghazvinian, R., Do, R., Thiru, P., Vergilio, J.-A., Beggs, A.H., Sieff,

C.A., Orkin, S.H., Nathan, D.G., Lander, E.S., others, 2012. Exome sequencing identifies GATA1 mutations resulting in Diamond-Blackfan anemia. J. Clin. Invest. 122, 2439–2443.
Schieppati, A., Remuzzi, G., 2005. Chronic renal diseases as a public health problem: epidemiology, social, and economic implications. Kidney Int. 68, S7–S10.
Schwarz, J.M., Rödelsperger, C., Schuelke, M., Seelow, D., 2010. MutationTaster evaluates disease-causing potential of sequence alterations. Nat. Methods 7, 575–576.
Schwede, T., Kopp, J., Guex, N., Peitsch, M.C., 2003. SWISS-MODEL: an automated protein homology-modeling server. Nucleic Acids Res. 31, 3381–3385. doi:10.1093/nar/gkg520
Seaquist, E.R., Goetz, F.C., Rich, S., Barbosa, J., 1989. Familial clustering of diabetic kidney disease. N. Engl. J. Med. 320, 1161–1165.
Seidman, J.G., Seidman, C., 2001. The genetic basis for cardiomyopathy: from mutation identification to mechanistic paradigms. Cell 104, 557–567.
Seikaly, M.G., Salhab, N., Gipson, D., Yiu, V., Stablein, D., 2006. Stature in children with chronic kidney disease: analysis of NAPRTCS database. Pediatr. Nephrol. 21, 793–799.
Severance, C., 2012. Inventing PHP: Rasmus Lerdorf. Computer 45, 0006–7.
Shendure, J., Ji, H., 2008. Next-generation DNA sequencing. Nat. Biotechnol. 26, 1135–1145.
Singh, S.K., Malik, A., Firoz, A., Jha, V., 2012. CDKD: a clinical database of kidney diseases. BMC Nephrol. 13, 23.
Soden, S.E., Saunders, C.J., Willig, L.K., Farrow, E.G., Smith, L.D., Petrikin, J.E., LePichon, J.-B., Miller, N.A., Thiffault, I., Dinwiddie, D.L., others, 2014. Effectiveness of exome and genome sequencing guided by acuity of illness for diagnosis of neurodevelopmental disorders. Sci. Transl. Med. 6, 265ra168–265ra168.
Soylemezoglu, O., Wild, G., Dalley, A.J., MacNeil, S., Milford-Ward, A., Brown, C.B., Nahas, A.M. El, 1997. Urinary and serum type III collagen: markers of renal fibrosis. Nephrol. Dial. Transplant. 12, 1883–1889.
Stanifer, J.W., Jing, B., Tolan, S., Helmke, N., Mukerjee, R., Naicker, S., Patel, U., 2014. The epidemiology of chronic kidney disease in sub-Saharan Africa: a systematic review and meta-analysis. Lancet Glob. Health 2, e174–e181.
Stenson, P.D., Mort, M., Ball, E.V., Shaw, K., Phillips, A.D., Cooper, D.N., 2014. The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Hum. Genet. 133, 1–9.
Stevens, P.E., Levin, A., 2013. Evaluation and management of chronic kidney disease: synopsis of the kidney disease: improving global outcomes 2012 clinical practice guideline. Ann. Intern. Med. 158, 825–830.
Stitziel, N.O., Fouchier, S.W., Sjouke, B., Peloso, G.M., Moscoso, A.M., Auer, P.L., Goel, A., Gigante, B., Barnes, T.A., Melander, O., others, 2013. Exome sequencing and directed clinical phenotyping diagnose cholesterol ester storage disease presenting as autosomal recessive hypercholesterolemia. Arterioscler. Thromb. Vasc. Biol. 33, 2909–2914.
Strathdee, C.A., Gavish, H., Shannon, W.R., Buchwald, M., 1992. Cloning of cDNAs for Fanconi's anaemia by functional complementation. Nature 356, 763–767.
Sumaili, E.K., Krzesinski, J.-M., Zinga, C.V., Cohen, E.P., Delanaye, P., Munyanga, S.M., Nseka, N.M., 2009. Prevalence of chronic kidney disease in Kinshasa:

results of a pilot study from the Democratic Republic of Congo. Nephrol. Dial. Transplant. 24, 117–122.
Swaminathan, S., Leung, N., Lager, D.J., Melton, L.J., Bergstralh, E.J., Rohlinger, A., Fervenza, F.C., 2006. Changing incidence of glomerular disease in Olmsted County, Minnesota: a 30-year renal biopsy study. Clin. J. Am. Soc. Nephrol. 1, 483–487.
Tan, R., Wang, Y., Kleinstein, S.E., Liu, Y., Zhu, X., Guo, H., Jiang, Q., Allen, A.S., Zhu, M., 2014. An evaluation of copy number variation detection tools from whole-exome sequencing data. Hum. Mutat. 35, 899–907. doi:10.1002/humu.22537
Taylor, J.C., Martin, H.C., Lise, S., Broxholme, J., Cazier, J.-B., Rimmer, A., Kanapin, A., Lunter, G., Fiddy, S., Allan, C., others, 2015. Factors influencing success of clinical genome sequencing across a broad spectrum of disorders. Nat. Genet.
Tennessen, J.A., Bigham, A.W., O'Connor, T.D., Fu, W., Kenny, E.E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G., Kang, H.M., Jordan, D., Leal, S.M., Gabriel, S., Rieder, M.J., Abecasis, G., Altshuler, D., Nickerson, D.A., Boerwinkle, E., Sunyaev, S., Bustamante, C.D., Bamshad, M.J., Akey, J.M., Broad GO, Seattle GO, NHLBI Exome Sequencing Project, 2012. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science 337, 64–69. doi:10.1126/science.1219240
Teppo, A.-M., Törnroth, T., Honkanen, E., Grönhagen-Riska, C., 2003. Urinary amino-terminal propeptide of type III procollagen (PIIINP) as a marker of interstitial fibrosis in renal transplant recipients. Transplantation 75, 2113–2119.
Theis, J.L., Sharpe, K.M., Matsumoto, M.E., Chai, H.S., Nair, A.A., Theis, J.D., de Andrade, M., Wieben, E.D., Michels, V.V., Olson, T.M., 2011. Homozygosity mapping and exome sequencing reveal GATAD1 mutation in autosomal recessive dilated cardiomyopathy. Circ. Cardiovasc. Genet. 4, 585–594.
Tibben, A., 2007. Predictive testing for Huntington's disease. Brain Res. Bull. 72, 165–171.
Tiffin, N., Hodkinson, B., Okpechi, I., 2013. Lupus in Africa: can we dispel the myths and face the challenges? Lupus 0961203313509296.
(US), N.C. for H.S., 1994. Plan and operation of the third National Health and Nutrition Examination Survey, 1988-94. Natl Ctr for Health Statistics.
Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., others, 2001. The sequence of the human genome. science 291, 1304–1351.
Veriava, Y., Toit, E. Du, Lawley, C.G., Milne, F.J., Reinach, S.G., 1990. Hypertension as a cause of end-stage renal failure in South Africa. J. Hum. Hypertens. 4, 379–383.
Via García, M., Consortium, 1000 Genomes Project, others, 2012. An integrated map of genetic variation from 1,092 human genomes. Nat. 2012 Vol 491 P 56-65.
Villa-Blanco, I., Calvo-Alén, J., 2012. Utilizing registries in systemic lupus erythematosus clinical research.
Vleming, L.J., Bruijn, J.A., Van Es, L.A., 1999. The pathogenesis of progressive renal failure. Neth. J. Med. 54, 114–128.
Wang, K., Li, M., Hakonarson, H., 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 38, e164–e164.

Watnick, S., 2007. Pregnancy and contraceptive counseling of women with chronic kidney disease and kidney transplants. Adv. Chronic Kidney Dis. 14, 126–131.
Wei, Z., Wang, W., Hu, P., Lyon, G.J., Hakonarson, H., 2011. SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data. Nucleic Acids Res. 39, e132–e132.
Wells, Q.S., Becker, J.R., Su, Y.R., Mosley, J.D., Weeke, P., D'Aoust, L., Ausborn, N.L., Ramirez, A.H., Pfotenhauer, J.P., Naftilan, A.J., others, 2013. Whole exome sequencing identifies a causal RBM20 mutation in a large pedigree with familial dilated cardiomyopathy. Circ. Cardiovasc. Genet. CIRCGENETICS–112.
Weng, L., Kavaslar, N., Ustaszewska, A., Doelle, H., Schackwitz, W., Hébert, S., Cohen, J.C., McPherson, R., Pennacchio, L.A., 2005. Lack of MEF2A mutations in coronary artery disease. J. Clin. Invest. 115, 1016–1020. doi:10.1172/JCI24186
Wetterstrand, K.A., 2013. DNA sequencing costs: data from the NHGRI Genome Sequencing Program (GSP). Natl. Hum. Genome Res. Inst.
Wheeler, D.A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.-J., Makhijani, V., Roth, G.T., others, 2008. The complete genome of an individual by massively parallel DNA sequencing. nature 452, 872–876.
Woo, H.-M., Park, H.-J., Baek, J.-I., Park, M.-H., Kim, U.-K., Sagong, B., Koo, S.K., 2013. Whole-exome sequencing identifies MYO15A mutations as a cause of autosomal recessive nonsyndromic hearing loss in Korean families. BMC Med. Genet. 14, 72.
Wooster, R., Bignell, G., Lancaster, J., Swift, S., Seal, S., Mangion, J., Collins, N., Gregory, S., Gumbs, C., Micklem, G., others, 1995. Identification of the breast cancer susceptibility gene BRCA2. Nature 378, 789–792.
Worthey, E.A., Mayer, A.N., Syverson, G.D., Helbling, D., Bonacci, B.B., Decker, B., Serpe, J.M., Dasu, T., Tschannen, M.R., Veith, R.L., others, 2010. Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. Genet. Med. 13, 255–262.
Wu, J., Jiang, R., 2013. Prediction of deleterious nonsynonymous single-nucleotide polymorphism for human diseases. ScientificWorldJournal 2013, 675851. doi:10.1155/2013/675851
Xiu, X., Yuan, J., Deng, X., Xiao, J., Xu, H., Zeng, Z., Guan, L., Xu, F., Deng, S., 2014. A Novel COL4A5 Mutation Identified in a Chinese Han Family Using Exome Sequencing. BioMed Res. Int. 2014.
Xue, Y., Chen, Y., Ayub, Q., Huang, N., Ball, E.V., Mort, M., Phillips, A.D., Shaw, K., Stenson, P.D., Cooper, D.N., Tyler-Smith, C., 2012. Deleterious- and Disease-Allele Prevalence in Healthy Individuals: Insights from Current Predictions, Mutation Databases, and Population-Scale Resequencing. Am. J. Hum. Genet. 91, 1022–1032. doi:10.1016/j.ajhg.2012.10.015
Yach, D., Hawkes, C., Gould, C.L., Hofman, K.J., 2004. The global burden of chronic diseases: overcoming impediments to prevention and control. Jama 291, 2616–2622.
Yandell, M., Huff, C., Hu, H., Singleton, M., Moore, B., Xing, J., Jorde, L.B., Reese, M.G., 2011. A probabilistic disease-gene finder for personal genomes. Genome Res. 21, 1529–1542.
Yang, Y., Muzny, D.M., Reid, J.G., Bainbridge, M.N., Willis, A., Ward, P.A., Braxton, A., Beuten, J., Xia, F., Niu, Z., others, 2013. Clinical whole-exome sequencing

for the diagnosis of mendelian disorders. N. Engl. J. Med. 369, 1502–1511.
Yan, X.-J., Xu, J., Gu, Z.-H., Pan, C.-M., Lu, G., Shen, Y., Shi, J.-Y., Zhu, Y.-M., Tang, L., Zhang, X.-W., others, 2011. Exome sequencing identifies somatic mutations of DNA methyltransferase gene DNMT3A in acute monocytic leukemia. Nat. Genet. 43, 309–315.
Yao, X.D., Chen, X., Huang, G.Y., Yu, Y.T., Xu, S.T., Hu, Y.L., Wang, Q.W., Chen, H.P., Zeng, C.H., Ji, D.X., others, 2012. Challenge in pathologic diagnosis of Alport syndrome: evidence from correction of previous misdiagnosis. Orphanet J Rare Dis 7, 100.
Yoon, S., Xuan, Z., Makarov, V., Ye, K., Sebat, J., 2009. Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Res. 19, 1586–1592.
Yu, H., Anderson, P.J., Freedman, B.I., Rich, S.S., Bowden, D.W., 2000. Genomic structure of the human plasma prekallikrein gene, identification of allelic variants, and analysis in end-stage renal disease. Genomics 69, 225–234.
Zhao, Y., Zhao, F., Zong, L., Zhang, P., Guan, L., Zhang, J., Wang, D., Wang, J., Chai, W., Lan, L., others, 2013. Exome sequencing and linkage analysis identified Tenascin-C (TNC) as a novel causative gene in nonsyndromic hearing loss. PloS One 8, e69549.