Université Fédérale
Toulouse Midi-Pyrénées

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par :

Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier)

Présentée et soutenue par :

Daniel Borràs Morales

le vendredi 13 octobre 2017

Titre :

The use of sequencing technologies for enhanced understanding of molecular determinants in renal diseases

---

École doctorale et discipline ou spécialité :
ED BSB : Bio-informatique, génomique et biologie des systèmes

Unité de recherche :
INSERM U1048/I2MC-Équipe 12

Directeur/trice(s)s de Thèse :
Dr. Joost Schanstra
Dr. Julie Klein

Rapporteurs :
Pr. Gerjan Navis (Rapporteur)
Pr. Benito Yard (Rapporteur)
Pr. Fulvio Magni (Rapporteur)

Membre(s) du Jury :
Pr. Isabelle Castan-Laurel (Présidente jury)
Pr. Gerjan Navis (Rapporteur)
Pr. Benito Yard (Rapporteur)
Dr. Joost Schanstra (Directeur)
Pr. Bart Janssen (Membre invité)

The use of sequencing technologies for enhanced understanding of molecular determinants in renal diseases.

L'utilisation de technologies de séquençage pour une meilleure compréhension des déterminants moléculaires dans les maladies rénales.

Daniel M. Borràs

Joost P. Schanstra (Director) Bart Janssen (Supervisor)

Index

Chapter 1

1. Introduction

Parts of this chapter were adapted from Daniel M. Borràs[1] and Bart Janssen[1] - "The use of transcriptomics in clinical applications." (under revision).

[1]*GenomeScan B.V., Pleasmanlaan 1d, Leiden*

1.1. Renal disorders in the health care system

Renal diseases have a great impact in the health care economy of any modern society. Among these renal diseases, Chronic Kidney Disease (CKD) has a global prevalence which was estimated from literature searches to be up to 13% (Table 1) (Hill et al., 2016). The lack of an early diagnosis of CKD and the asymptomatic nature of CKD in early stages, hampers the accurate calculation of the real prevalence, which may be even higher than the estimated 13% (Hill et al., 2016). In addition to the high economical cost associated to CKD, in all of its different stages, CKD is an independent risk factor for cardiovascular disease (CVD) which is a primary cause of morbidity and mortality (Go et al., 2004; Hill et al., 2016). Advanced CKD stages, namely stages 4 and 5, are associated with low kidney function that result in severe clinical manifestations. Advanced CKD stages can progress to end stage renal disease (ESRD) in which there is a complete loss of kidney function, and is associated with even higher healthcare costs of maintaining a dialysis program or renal replacement therapy (RRT) (Wetmore and Collins, 2016).

Table 1: Prevalence of CKD for stages from 1 to 5 in USA, Canada and Europe represented as mean prevalence and confidence interval extracted from 100 CKD studies from year 2000 (Hill et al., 2016).

	CKD stages 1 to 5	**CKD stages 3 to 5**
	Prevalence (%)	*Prevalence (%)*
USA, Canada	15.45 (11.71, 19.20)	14.44 (8.52, 20.36)
Europe	18.38 (11.57, 25.20)	11.86 (9.93, 13.79)

**Adapted from Hill et al, 2016.*

Table 2: Most common causes of ESRD, USA (www.usrds.org).

Cause	**ESRD patients (%)**
Diabetes mellitus	45
Hypertension	27
Glomerulonephritis	8
Polycystic kidney disease	2
Other/Unkown	18

**Adapted from Hall and Guyton, 2011.*

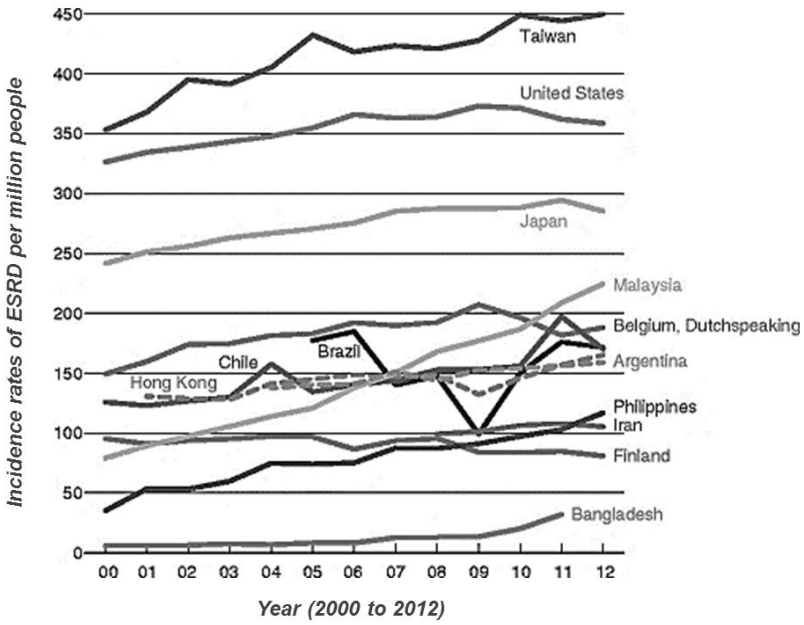


Figure 1: Incidence rates of ESRD, by country, from 2000 to 2012.

**Adapted from Wetmore and Collins, 2016.*

Since most renal disorders evolve into ESRD in their final stages, a global measure of the impact of renal diseases in the healthcare costs could be abstracted as the number of cases of ESRD per year. Using these measurements, we can observe that there have been advancements in the overall health outcomes related to kidney diseases over the past decade (Figure 1). The incidence rates of ESRD seem to be stabilized for some countries in recent years despite an overall increasing trend in ESRD cases (Figure 1) (Wetmore and Collins, 2016). However, these trends do not represent the health care costs of each country which can vary case by case depending on patient accessibility to dialysis and Renal Replacement Therapy (RRT) among other factors (Wetmore and Collins, 2016). The increased survival of 20% observed for dialysis patients in USA from 2003 to 2012 may explains the growth in the prevalent population of dialysis patients despite the opposite trend observed in incidence rates of ESRD in developed countries (Figure 1) (Wetmore and Collins, 2016). Among the most common cause of ESRD there are glomerulonephritis and polycystic kidney disease, as well as diabetes mellitus and hypertension (Table 2) (Hall and Guyton, 2011). Therefore, it is important to stress the relevance of any research performed in the context of renal diseases, which will have a great impact in terms of improvement of the healthcare costs associated to renal disorders, and most important the quality of life as well as life expectancy of patients suffering from kidney diseases.

1.2. Understanding the renal system, anatomy and physiology

The kidneys are two bean-shaped organs, located in the abdominal cavity, which have the important function of eliminating toxic and waste products from the blood stream. Waste products include direct or side products of metabolic processes that are no longer needed, as well as some exogenous chemicals, drugs. The most important waste products are urea resulting from amino-acid metabolic processes, creatinine from muscle fiber metabolism, uric acid from metabolic processes involving nucleic acids, hemoglobin degradation products such as bilirubin, and side metabolites produced by hormone processing metabolism. Other kidney functions of major importance include maintenance of fluid and electrolyte homeostasis as well as blood pressure, and excretion/metabolism of hormones amongst other

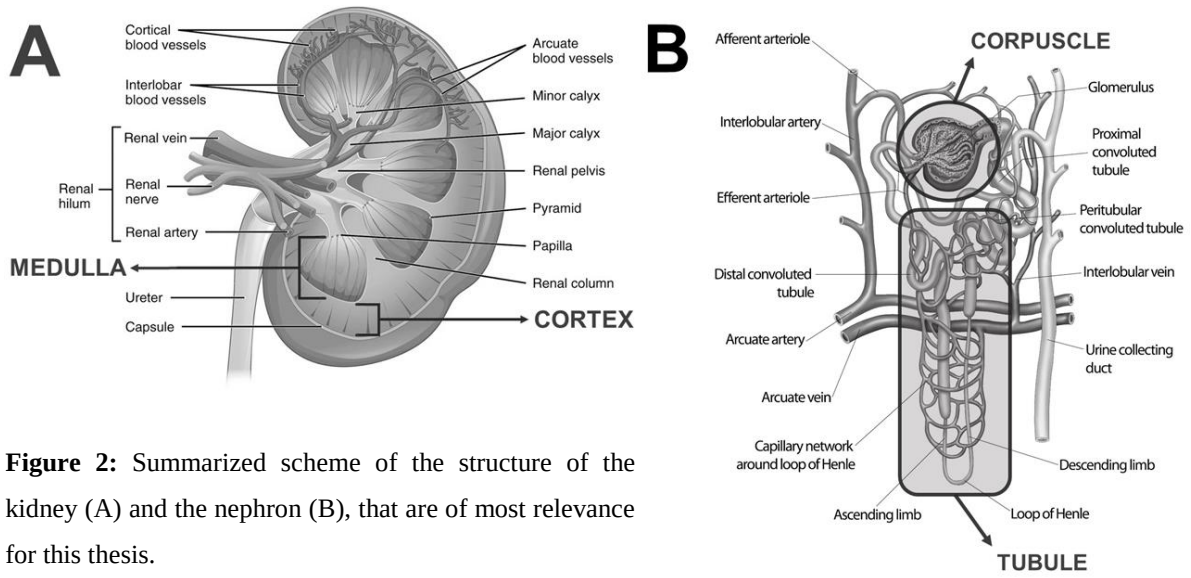


Figure 2: Summarized scheme of the structure of the kidney (A) and the nephron (B), that are of most relevance for this thesis.

metabolic functions such as acid-base balance and gluconeogenesis. Most of these functions performed by the kidney occur in the outer cortex or the inner medulla (Figure 2). Each human kidney contains large numbers of nephrons (≥200,000 up to ≤2 million) (Denic et al., 2017), that are the functional unit of the kidneys. The nephron, can be divided into two main units, a: the corpuscle, including the glomerulus and Bowman's capsule where large amounts of plasma are filtered (Figure 2); and b: the tubule, which concentrates and processes the excreted fluids to be converted into urine (Figure 2).

One of the primary aspects of excretion is the regulation of water and electrolyte balances, directly adjusted by their excretion levels. These, combined with the secretion of vasoactive substances such as renin, takes also part in the kidney's function to regulate arterial pressure. The difference in hydrostatic pressure and osmotic pressure between the glomerular vessels

and the Bowman's space will determine the fluid movement across the capillary walls and thereby the production of ultrafiltrate plasma (primary urine). In the course of 24 hours, around 180 liters of plasma is normally filtered by the glomeruli. Under physiological conditions, the glomerular filtration barrier should retain cellular elements, and most of proteins from the ultrafiltrate. Substances that pass this filtration are further reabsorbed by the tubular system recovering the majority of wanted substances such as water, small proteins and electrolytes. The remaining substances will then form part of the excreted urine (final urine) with the water that was not reabsorbed. The efficiency of filtration performed by the kidney is also known as glomerular filtration rate (GFR) which can be measured by the creatinine levels observed in blood and the urine. Creatinine is a side product of muscle metabolism that is freely filtered by the glomerular barrier, but not re-absorbed by the nephron. Accordingly, excreted creatinine per minute represent a direct measure of the creatinine that has been filtered and provides the means to measure the GFR (ml/min).

Table 3: Hormones that regulate the Glomerular Filtration Rate.

		GFR
Vasoconstrictors	*Angiotensin II*	↘
	Endothelin	↘
	Sympathetic nerves	↘
Vasodilators	*Prostaglandins*	→ or ↗
	Nitric oxide	↗
	Bradykinin	↗
	Natriuretic peptides	↗

**Adapted from Koeppen and Stanton, 2013.*

Monitoring GFR in the clinic can indicate disease states which often affect the filtration surface area decreasing the ultrafiltrate. GFR may change because of different factors associated with renal diseases such as reduction of functional glomeruli, constriction of glomerular arterioles, decreased renal perfusion either by increased afferent or decreased efferent arteriole resistance, reduced glomerular oncotic pressure of proteins, and increased Bowman's space hydrostatic pressure by obstruction. In addition, there are many hormones that can influence the GFR such as angiotensin, and prostaglandins (Table 3).

Overall, the renal system is a crucial part of the human body. Renal diseases that impede the proper functioning of the kidneys can have a greater impact than the localized damage because of its main roles on the excretion of waste products, chemicals, drugs and hormone metabolites, the regulation of water and electrolytes, the regulation of blood pressure, the buffering of blood acid-base balance, and the regulation of erythrocyte and vitamin D production among other indirect functions. Therefore, the understanding of kidney-damaging mechanisms and the recovery of kidney function are key for the protection of patients from further advancing into more severe stages.

1.3. Renal disorders

Severe kidney diseases may be grouped into to two main categories depending on the period that they occur, (*1.*) acute renal failure, also known as acute kidney injury (AKI) which involves an abrupt stop of the normal functions of the kidney; and (*2.*) chronic renal failure,

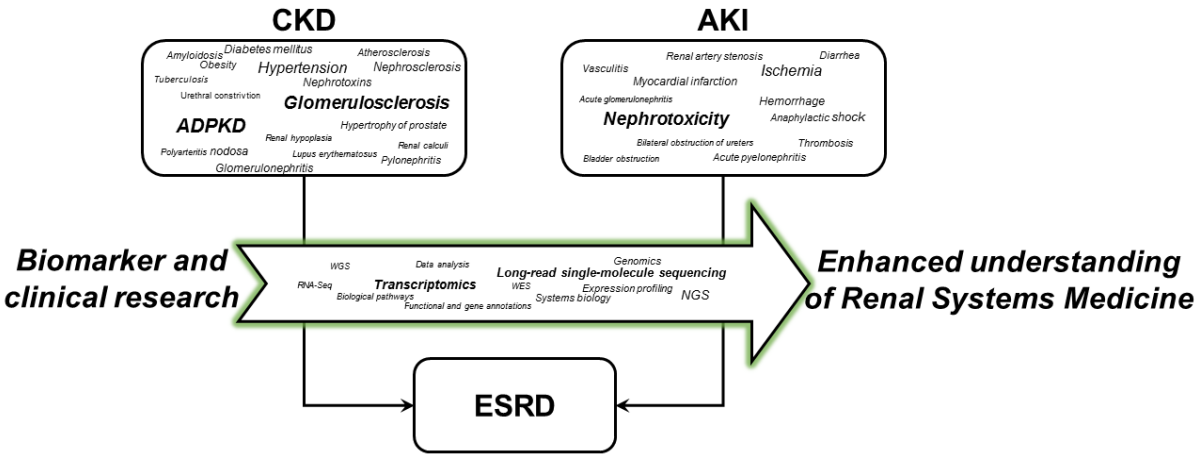


Figure 3: Scheme of common tools and research approaches used for biomarker and clinical research applied to renal disorders with a common endpoint, ESRD. Highlighted (**bold**) are key diseases, concepts and terms that are addressed in this thesis.

also known as CKD, which differs from AKI by its gradual and permanent loss of renal functions. We believe important to clarify at this point that this thesis is based on integrative omics approaches. This includes technical and methodological approaches such as NGS, data analysis, as well as systems biology, and renal systems medicine methodologies. Renal disorders are very diverse with many etiologies that converge into a common outcome, ESRD (Figure 3). Hence, we focused the background information provided in this introduction

chapter to 3 selected renal diseases that are further studied in this thesis such as ADPKD, Glomerulosclerosis, and AKI.

1.3.1. Chronic Kidney Disease

Chronic renal failure, also known as CKD, is mainly defined by an irreversible loss of kidney function that has its origin in the decrease of functional or healthy nephrons. Nephrogenesis is generally assumed to be completed few weeks before birth, when the full set of nephrons should be complete and will gradually reduce in number along the years (Hoy et al., 2010; Denic et al., 2017). The renal system, despite its important role for the proper functioning of the human body, cannot regenerate new nephrons. The permanent loss of renal functional units could be estimated relative to the number of functional glomeruli in the kidney, and showed a direct relationship of functional units with the decrease of GFR (Denic et al., 2017). It is well established that any injuries that the kidneys suffer, such as damage produced by aging, physical injuries, or disease-associated injuries, would accumulate during the years and have a strong impact on the proper functioning of the kidneys. This cumulative effect would lead to a vicious cycle of continuous deterioration and progression of loss of kidney function that would evolve into CKD and ESRD in the more severe cases (Hall and

Table 4: Values of excretion of the overall nephrons and per nephron in renal failure.

		Normal kidneys	**75% Nephron-loss kidneys**	
All nephrons	*Number of nephrons*	2	0.5	Millions
	GFR	125	40	$^{ml}/_{min}$
	Volume excreted	1.5	1.5	
Single nephron	*GFR*	62.5	80	$^{nl}/_{min}$
	Volume excreted	0.75	3	

**Adapted from Hall and Guyton, 2011.*

Guyton, 2011). The majority of CKD cases passes unnoticed until the number of functional nephrons is drastically reduced. Normal kidney function can usually be maintained until the number of functional nephrons drop below 20 or 25% of normal amounts when severe clinical symptoms will appear (Table 4). The main causes of nephron loss that progress to CKD may include a wide variety of disorders including metabolic disorders (diabetes, obesity), hypertension, immunologic disorders (glomerulonephritis, lupus), vascular disorders (atherosclerosis), infections (tuberculosis), primary tubular disorders (nephrotoxins), and urinary obstructions (renal calculi) among others (Table 5).

Table 5: Some of the main causes of CKD

Metabolic Disorders	*Diabetes mellitus* *Obesity* *Amyloidosis*
Hypertension	
Renal vascular disorders	*Atherosclerosis* *Nephrosclerosis-hypertension*
Immunologic disorders	*Glomerulonephritis* *Polyarteritis nodosa* *Lupus erythernotosus*
Infections	*Pyelonephritis* *Tuberculosis*
Primary tubular disorders	*Nephrotoxins (analgesics, heavy metals)*
Urinary tract obstructions	*Renal calculi* *Hypertrophy of prostate* *Urethral constriction*
Congenital disorders	*Polycystic disease* *Renal hypoplasia*
**Adapted from Hall and Guyton, 2011.*	

1.3.1.1. Autosomal Dominant Polycystic Kidney Disease (ADPKD)

ADPKD is a common inherited kidney disease that accounts for 5% to 10% of ESRD cases (Spithoven et al., 2014). Most causing mutations of ADPKD occur in *PKD1* and *PKD2* genes with a reported prevalence of 75-85% and 15-25%, respectively (Barua et al., 2009; Harris and Rossetti, 2010). This renal disease phenotype is characterized by a phenotype of enlarged kidneys because of the presence of numerous renal cysts. Since ADPKD is a multisystem disorder, cysts are not solely limited to the renal tissue but may appear in other organs as well such as the liver, seminal vesicles or the pancreas among others. On average, at the age of 58-60, 50-70% of the ADPKD patients progress to ESRD (Harris and Torres, 1993; Gansevoort et al., 2016).

Mutations in *PKD1* and *PKD2*, may disrupt the functions of these genes and are the main cause of ADPKD. *PKD1* (Polycystin 1) and *PKD2* (Polycystin 2) are two transmembrane proteins expressed in the primary cilia that are involved in the reabsorption of Ca^{2+}. Activation of *PKD2* by *PKD1* allows the flow of Ca^{2+} increasing the intracellular Ca^{2+}, which in return also activates K^+ channels to secrete K^+ into the tubular fluid. Disruptions in the amino-acid sequence of *PKD1*, in the form of DNA variants, not only affects the Ca^{2+} uptake because of its interaction with *PKD2*, but will also alter *PKD1*'s signaling pathways,

including kidney cell proliferation, cell adhesion, cell migration, cell differentiation, and apoptosis (Koeppen and Stanton, 2013; Castelli et al., 2015). Because of the close interaction between *PKD1* and *PKD2*, mutations in any of these genes will play a role in the cysts formation and the development of ADPKD, particularly with truncating mutations in *PKD1* (Gainullin et al., 2015; Gansevoort et al., 2016).

The current treatment strategy for ADPKD include the use of vasopressin V2 receptor antagonist (Tolvaptan) to slow the advance of ADPKD patients with CKD in stages from 1 to 3 that show clear signals of rapid progression (Yu et al., 2015; Gansevoort et al., 2016). However, due to the side effects of this drug, and its cost, a clear selection of patients is needed to identify patients that could benefit the most from this treatment. Promising results were obtained with mTOR inhibitors such as rapamycin using animal models, but the same efficacy could not be observed on human clinical trials (Yu et al., 2015). Other research on therapeutic compounds for ADPKD have showed that *STAT3* inhibitors such as pyrimethamine, S3I-201, or the broad spectrum diferulomethane (curcumin) have great potential to become a new generation of drugs for the treatment of ADPKD (Leonhard et al., 2011; Harris and Torres, 2014).

The diagnosis of ADPKD by screening of *PKD1* is still challenging due to its high homology with six other *PKD1* pseudogenes (Rossetti et al., 2007). There have been several attempts to use next generation sequencing (NGS)-based diagnostic approaches to improve *PKD1* resolution for diagnostic applications (Tan et al., 2014; Trujillano et al., 2014; Eisenberger et al., 2015; Mallawaarachchi et al., 2016). These approaches rely on the high-quality sequence information that can be obtained from Illumina sequencing platforms at relatively low cost (Su et al., 2011; Mardis, 2013; Oliver et al., 2015). However, short-read NGS sequencing approaches such as whole genome (WGS) and whole exome sequencing (WES), often fail at reliably characterizing complex regions of the human genome (Lee and Schatz, 2012; Chaisson et al., 2015). These regions are quite often associated with high GC-content, segmental duplications (SDs), low complexity sequences and gaps still present in the reference sequence of the human genome (Lee and Schatz, 2012; Steinberg et al., 2014; Chaisson et al., 2015). Some of these complex regions could be resolved using single-molecule long-read sequencing, which can improve our understanding of genetic variations in complex but clinically relevant genomic regions (Guo et al., 2013; Loomis et al., 2013; Laver et al., 2016; Qiao et al., 2016). Therefore, we used single-molecule long-read sequencing as a

tool with added value to characterize genetic variants from complex genomic regions such as *PKD1*, which serves as an excellent example of a challenging and complex locus.

1.3.1.2. Glomerulosclerosis

Different types of sclerosis can lead to ischemia and damage the kidney tissue including atherosclerosis, nephrosclerosis and glomerulosclerosis. Naturally occurring sclerosis occurs and accumulates with age (Denic et al., 2017). Typically, after a small lesion in the vessel, fibrinoid deposits accumulate in response to small plasma leakage. This cause the vessel walls to become thicker, eventually ending in a permanent occlusion. If this occur within the glomeruli, the injury is named glomerulosclerosis. The loss of functional glomeruli decreases with age while the number of sclerotic glomeruli increases, with an estimated 50% loss of total nephrons from the age of 18-20 to 70-75 years (Denic et al., 2017). There is a trend of increased loss of nephrons with nephrosclerosis and glomerulosclerosis, as well as hypertension (Denic et al., 2017). Other causes of glomerulosclerosis include the most common form of glomerular disease that leads to ESRD, focal segmental glomerulosclerosis (FSGS) (Kiffel et al., 2011). Up to 80% of FSGS patients were identified within the group of idiopathic nephrotic syndrome patients that not respond to steroid therapies (Han and Kim, 2016). There seems not to be a clear alternative treatment for FSGS patients that are corticosteroid resistant, and the many of these progress to ESRD (Han and Kim, 2016). Some second line alternatives provided mixed results due to effectiveness, nephrotoxicity, or other adverse reactions depending on each individual case such as cycloshosphoamide, cyclosporine, rituximab, abatacept, adalimumab, or fresolimumab (Kiffel et al., 2011; Han and Kim, 2016). Regardless of the source cause of FSGS, renal fibrosis and loss of kidney function will progress in patients that are resistant to therapies. Further studies to understand the underlying mechanisms of fibrosis may aid the development of alternative therapies to diminish the progression of nephrosclerosis and glomerulosclerosis.

1.3.2. Acute kidney injury

Main causes of AKI are included in three categories defined as prerenal, intrarenal, and postrenal AKI. Prerenal causes include abnormalities that originate outside the kidneys such as heart failure or haemorrhages that produce low blood pressure, and volume (Table 6). The intrarenal AKI results from internal abnormalities from the kidney that may affect the blood vessels from the kidney, the glomeruli or the tubules, and postrenal AKI is caused by abnormalities in the lower urinary tract, usually blocking the urine flow (Table 6). There exist

no specific therapies that can attenuate AKI and improve the recovery of the kidney function after an AKI insult (Bellomo et al., 2012). Therefore, current treatments are mostly supportive or preventive, including dialysis, with the intention of reducing the loss of kidney function and prevent the progression into CKD, or even ESRD (Bellomo et al., 2012; Harty, 2014). AKI episodes are more frequent in hospitalized patients under surgery or critically ill as well as patients following high risk drug treatments such as chemotherapies (Bellomo et al., 2012). AKI episodes pose a substantial risk to these patients, including increased morbidity, progression to severe CKD stages, or death.

Exposure to some chemicals or drugs such as antibiotics, immunosuppressants, cancer chemotherapy, non-steroidal anti-inflammatory drugs among others, have an undesired nephrotoxic effect that is considered among the most common causes of AKI observed in humans, alongside ischemia, and urinary obstructions (Yang et al., 2010). A common and accepted model of nephrotoxicity is folic acid nephropathy which was also previously reported to occur in humans (Metz-Kurschel et al., 1990). This causes reversible increase in serum creatinine and urea, tubular cell death, compensatory tubular cell proliferation,

Table 6: List of some of the causes of AKI.

Prerenal	Intravascular volume depletion	*Hemorrhage, Diarrhea, vomiting.*
	Cardiac failure	*Myocardial infarction, valvular damage.*
	Peripheral vasodilation and resultant hypotension	*Anaphylactic shock, anesthesia, sepsis and severe infections, primary renal hemodynamic abnormalities, renal artery stenosis, and embolism or thrombosis of renal artery or veins.*
Intrarenal	Small vessel and glomerular injury	*Vasculitis (polyarteritis nodosa), cholesterol emboli, malignant hypertension, acute glomerulonephritis.*
	Tubular epithelial injury (tubular necrosis)	*Acute tubular necrosis due to ischemia, acute necrosis due to toxins (heavy metals, ethylene glycol, insecticides, poison mushrooms, carbon tetrachloride).*
	Renal interstitial injury	*Acute pyelonephritis, acute allergic interstitial nephritis.*
Postrenal		*Bilateral obstruction of ureters (large stones or blood clots), bladder obstruction, urethra obstruction.*

**Adapted from Hall and Guyton, 2011.*

activation of an inflammatory response, and eventual progression to mild fibrosis (Fang et al., 2005; Doi et al., 2006; Ortega et al., 2006). Previous studies performed using mouse models showed that *Gdf15* gene may have a protective effect against tubular and interstitial damage (Mazagova et al., 2013). The role of *Gdf15* and the mechanisms of recovery of the kidney function are still poorly understood (Breit et al., 2012; Mazagova et al., 2013; Vallon, 2016).

Therefore, within this thesis we used an unbiased RNA sequencing approach to unravel the genetic drivers of renal tissue protection and recovery led by *Gdf15* expression.

1.4. Sequencing approaches in clinical applications

The currently available, and rapid advancement of sequencing technologies allows for the use of nucleic-acid material for clinical use with single base precision. Screening for multiple disease determinants, obtaining information on disease causes or determining a potential response to a particular treatment is now possible in a single run or test. Despite this, sequencing-based screening approaches are still not preferred over PCR tests mainly because of price. However, in complex diseases where multiple genes are implicated, such as acute myeloid leukemia, breast cancer, cardiovascular diseases, and some renal diseases, we can find good examples of the added diagnostic value of sequencing-based methods compared to traditional approaches. As discussed above, we focused our research efforts to study kidney diseases that could benefit from new advancements in NGS technologies and data analysis approaches such as ADPKD, AKI, or glomerulosclerosis.

The associations between genes and diseases have already been the subject of study for many decades. The clearest cases of associations showed that a particular change in a single gene can be the potential cause for a particular disease. There are over 1500 defined genes that were classified as monogenic disorders with an associated phenotype (Brinkman et al., 2006), but these do not cover most of the human diseases which are mainly multi-factorial. In their expert opinion Stylianos Antonarakis and Jacques Beckmann state that monogenic disorders are an unfortunate casualty in the race to find the determinants of complex diseases (Antonarakis and Beckmann, 2006; Brinkman et al., 2006; O'Connor and Crystal, 2006). Not all genetic mutations are detected or diagnosed using the same type of material. Isolated DNA is the most common type of patient material used for diagnostics applied to nucleotides. However, the analysis of other nucleic acids such as messenger RNAs (mRNAs) has also great potential to elucidate many genetic disorders. In particular, RNA can be used to diagnose complex diseases where multiple genes are implicated such as cardiovascular disease, breast cancer, and type 2 diabetes mellitus (Ross et al., 2008; Herder et al., 2011).

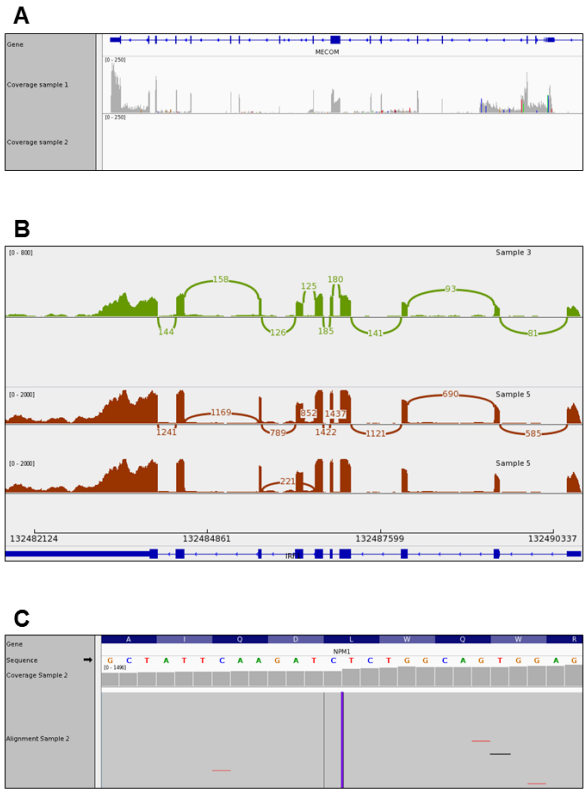


Figure 4: Visualization examples of a combined diagnostic test using RNA-seq to identify different types of nucleic factors in AML patients. (*A*) Gene coverage for RNA-Seq data. First gene track (blue line) shows the gene structure of *MECOM* by RefSeq gene annotation. Thick lines show exon locations and arrowed lines intronic locations. Second and third track show the coverage of *MECOM* for samples 1 and 2, respectively. In grey, the depth of coverage of sample 1 indicates that there is expression where the peaks are located. The extent of the expression is proportional to the height of the peaks, and the amplitude represents how much of the gene sequence is expressed, usually coinciding with exon boundaries. (*B*) Visual representation of splicing detected for two sequenced patients with RNA-Seq. Bottom blue track represents gene structure of *IRF1* gene. Thick lines show exons bases and arrowed line intronic bases. Green track shows the coverage of sample 3 (green blocks) as well as the number of reads that span the distance to the next splice acceptor region (green line). Red tracks show coverage of sample 5 (red blocks) and number of reads that span the distance to the next splice acceptor region (red line) connecting two exons. Last red track show an alternative splicing event with 221 reads that are connecting exons 3 and 5 but skipping exon 4. This shows that exon 4 is partially skipped by some transcripts in sample 5. (*C*) Visualization of a SNV detected by RNA-Seq. The 4bp insertion (TCTG) is represented as a purple line on the Alignment track for the *NPM1* gene. The gene track (blue) shows the codons present in this particular location (blue and dim blue) as well as the amino acid it encodes for (letters). The sequence of this particular location is shown and the black arrow gives information about the strand of the shown nucleotides (forward). Third track shows the coverage for each of the represented bases. Finally, the alignment is shown where horizontal grey lines represent a match with the reference sequence and a colored horizontal line represents a mismatch, which is colored with the same colors as the sequence nucleotides, being T red, A green, C blue, and G orange, black represents a deletion.

**(Primary data was kindly provided by Prof. Veelken, H., Head of Department of Hematology, Leiden University Medical Center)*

Over the years many diagnostic tests have been developed that took advantage of the latest advances in technology for biomarker discovery. The ultimate goal of molecular diagnostics is to accurately predict the presence or absence of a particular genetic disorder, infectious disease, or even response to a particular drug treatment. Towards this end, the latest advances in high throughput technologies opened up new opportunities for nucleic acid-based diagnostics allowing the development of new and more sensitive disease diagnostic tools. This goal is achieved by applying advanced statistical methods that reveal hidden patterns within the data that can characterize multi-factorial diseases. This was accomplished to some extent with the appearance of high throughput technologies such as micro-arrays and

sequencing technologies, both able to analyze thousands of measurements in the DNA or the RNA in a single run. The global analysis of the DNA mutations or mRNAs expression can be quite rewarding and informative, providing information on many variations with base pair accuracy in the case of some diagnostic tools. The analysis of not only the DNA variants but also the analysis of transcription products in diagnostics allowed to detect many diseases caused by genomic alterations as well as specific transcript modifications such as single nucleotide polymorphisms (SNP), small variants (SV), translocations, inversions, chimeric genes, breakpoints, post-transcriptional modifications, alternative splicing and gene expression (Dominissini et al., 2011; Su et al., 2011; Sánchez-Pla et al., 2012; Shyr and Liu, 2013; Mimura et al., 2014; Chaussabel, 2015). Many different methodologies commonly used for diagnosis of complex diseases are nowadays possible to combine in a single NGS test (Griffionen et al., 2016), such as the detection of gene expression, splicing variants, and SVs (Figure 4) in acute myeloid leukemia identified in *MECOM*, *IRF1*, and *NPM1*, respectively. Furthermore, global sequencing approaches can be used towards clinical diagnostics on autoimmune disorders, cancer, and infectious diseases (Chaussabel, 2015) as well as provide an alternative non-invasive procedure that could detect transplant rejection in early stages (Chen et al., 2013).

1.4.1. Prior to sequencing, sample material and preparation

In order to perform a diagnostic test on DNA, mRNA or any other source of nucleic acid material, the DNA or RNA first needs to be accurately extracted. There are many extraction methods available, and their applicability mainly depends on the sample source and type of material. Each method can provide specific advantages depending on the variety of accessible biological sample types and the downstream analysis to be performed. Some of these

advantages can be evaluated in several ways but yield and integrity are generally accepted as criteria of successful extraction. The material obtained from the sample extraction, regardless of the method used and the success of the extraction, is generally in the form of DNA or total RNA (Devonshire et al., 2013). In this state, the DNA isolated is usually quite pure and ready to be used for sequencing. However, the RNA extracted material from human samples contains mainly ribosomal RNA (rRNA) transcripts (approximately 80%), and tRNA transcripts (approximately 15%). The presence of these two RNA types greatly reduces the relative abundance of the other transcript species which are of diagnostic interest such as mRNAs. Therefore, an additional enrichment step of the preferred RNA species, such as mRNA, is needed after the extraction of total RNA. This enrichment of mRNA will further enhance the reliability and sensitivity of the diagnostic test.

The isolation of DNA or RNA from the biological samples can be currently achieved by several methods including magnetic bead-based extraction, silica columns and acid phenol/chloroform. This is usually achieved by several steps, which may include degradation steps of non-desired nucleic acids, amplification of DNA as well as hybridization and ligation amongst others. The degradation of the non-desired DNA carry over from an RNA sample can be performed by DNase treatments. The degradation of non-desired RNAs from a DNA sample can be performed by RNase degradation. Such RNA degradation is, for instance, performed after the reverse transcription amplification of enriched mRNAs by using poly-A primers. The amplification provides the first strand of the mRNAs, as cDNA, which will no longer be degraded by the RNase activity. Similarly, the enrichment can be achieved without RNase degradation by including adapters during the reverse transcription steps and/or ligation steps. Similar ligation steps can be performed for DNA enriched samples. Then the isolated DNA from a DNA sample or cDNA from an RNA sample is amplified by adding primers which are specific to the adapters introduced. There are other methods which involve physical separation of the mRNA transcripts. In this situation, the use of poly-A ligands is quite usual, also known as poly A capture enrichment. A common way of physical separation of nucleic acid material is by pulling down the DNA or RNA with magnetic beads. In the case of mRNA the poly-A ligand is attached to the beads and then the mRNA is hybridized and separated from the sample solution by the use of magnetic forces that pull the magnetic beads down. Another common method involves the use of filtration columns. In this situation, the nucleic acids are usually attached or captured to the stationary phase of the column by filtration or ligation (poly-A ligands for mRNA). In the majority of the described

enrichment cases, commercial kits may be available and used for the enrichment of DNA or RNA species such as mRNAs. It is important to note that some kits allow direct isolation of mRNAs without a previous step of total RNA extraction. As general recommendation, the evaluation of the performance of enrichment protocols and available kits is considered as good practice for the success of setting up any DNA- or RNA-based diagnostics test. Another method of enrichment is the polyacrylamide gel electrophoresis which is generally used to separate nucleic acid species of desired sequence length (Petrov et al., 2013).

One of the limitations of handling RNA, opposite to DNA, is that is quite prone to fast degradation. Unless the analysis of the RNA is performed immediately after the sampling process, effective precautions against degradation must be taken. Preventing RNA degradation is of particular concern for quantitative analyses and the diagnosis of diseases where high sensitivity is required. There exist several procedures and reagents that are used to preserve the integrity of RNA. In the majority of cases, the use of these measures is highly recommended. An extensively used method for preserving the RNA is to snap freeze the samples, which can be done with liquid nitrogen followed by storage for a longer period at -80C. Otherwise, during the collection of the samples, cells can be directly embedded in RNA safeguarding reagents or buffers which can preserve RNA from degradation. In this case, it may be possible to store the samples at a higher temperature depending on the approach and chemicals used during the collection. For all of the aforementioned cases, extra care has to be taken to maintain the samples at the lowest temperature possible during the handling and to work with appropriate RNase free solutions, material and working environment at all times (Devonshire et al., 2013).

Although it is always preferred to use “fresh” biological material for the analysis of nucleic acids, particularly for RNA, the use of fresh material is not always an option. In some cases, tissue samples are needed for clinical diagnostics which involve a particularly difficult and painful process for obtaining the material such as renal biopsies. Hence, it is not surprising that the current trend for the development of new diagnostics tools is focused on sample sources that are more easily obtained, in addition to the advantage of using non-invasive methods such as blood, saliva and urine (Devonshire et al., 2013).

When DNA or RNA is extracted from a biological sample for disease diagnostics we must also take into consideration that sample composition is another important variability introducing factor. Tissue samples, for instance, contain different cell types. Each cell type

will express its own gene repertoire or mutation profile. In the case of gene expression profiling, a sample's mRNA composition may also be influenced by internal, and external factors (i.e. nutrition, circadian stage, cellular cycle, stress, exercise or disease state). Sasagawa *et al.*, showed that single cell transcriptome analysis using RNA-Seq was able to identify and quantify non-genetic cellular heterogeneity, and even differentiate cell types and cell cycle phases of a single cell type (Sasagawa et al., 2013). Similarly, the accumulation of DNA mutations in clonal cancer cells, caused by the development of a cancer, can be quantified to reconstruct the phytogenic evolution and identify source or causal mutations (Bozic et al., 2016). Therefore, it is important to use appropriate methods for targeted cell type enrichment, if possible, such as laser capture micro dissection (LCM) for selecting tissue areas from tissue slides, cell sorting for enriching the cell fraction of interest, or centrifugation for separating the desired cell population (Todd and Kuo, 2002; Taussig et al., 2010; Devonshire et al., 2013; Sasagawa et al., 2013; Gutierrez-Arcelus et al., 2015).

Many of the aforementioned processes and techniques used for the extraction of nucleic acids are based on the use of a variety of solutions and chemicals. Unfortunately, the procedures are not perfect, which means that during the extraction there is a carry-over of chemicals and reminiscent sample components such as DNA (in the case of RNA isolation), proteins, and salts. Chemicals such as ethanol, chloroform or phenol, as well as monovalent cation salts such as ammonium acetate, lithium chloride and sodium acetate, may be present in the sample due to the extraction and precipitation approaches used (Walker and Lorsch, 2013). Carry-over genomic DNA may interfere with amplification or hybridization steps needed for RNA library preparations but can be easily removed from the sample by a DNase enzymatic treatment. Other remaining proteins can be readily removed by proteases. Sample stabilizers such as citrate, EDTA, and heparin may inhibit the reverse transcription (RT) and should also be removed. Several authors reported approaches for this purpose such as precipitating RNA with lithium chloride to get rid of sample heparin (Del Prete et al., 2007) or the use of ultracentrifugation (Ding et al., 2011). It is worth mentioning that any single effort towards the removal of potential contaminants may improve the outcome of the diagnostic test applied. These contaminants, in the end, may negatively influence the possible choices for downstream applications and in the worst case have a direct impact on the diagnostic assay and its outcome (Devonshire et al., 2013).

While keeping in mind that there are many possible sources of variability, one has to choose from a large series of isolation kits and methods. The choice should be made depending on the sample type to be analyzed. The ideal method is fast and easy, while allowing a reproducible extraction of nucleic acids that is immediately ready to use or store. Many of these kits already include cleaning steps from inhibitors and contaminants and could be automated. Additionally, in the case of RNA isolation, some of these kits may also directly extract not only total RNA but targeted subspecies of RNA such as mRNA, miRNA, and even viral RNAs, making the process much easier to set up for diagnostic applications. Several authors reviewed the performance of different isolation kits applied to a variety of tissue samples. A recent example of assessment of the performance of two purification kits was provided by Akutsu *et al.,* (2015). The authors compared RNeasy Mini Kit (silica column based kit from Qiagen) and EZ1 RNA Tissue Mini Kit (automatic magnetic beads based from Qiagen). The extraction was applied to two different biofluid samples, saliva and blood, in the form of fresh and dried cotton swap stains. The results of this comparison showed that silica column based extraction provided a slightly better RNA quality on fresh samples, and substantially better on sample stains. Another comparison between two isolation kits for DNA extraction applied to FFPE cell lung carcinoma tissue samples showed that while both kits were suitable for the identification of mutations despite one of the kits outperformed its competitor in yield and quality (Hu et al., 2014). These comparisons between multiple kits highlighted that there may be kit-dependent differences in acid nucleic extraction that may impact downstream mutation analysis and gene expression measurements (Hu et al., 2014; Javadi et al., 2014; Jeffries et al., 2014). Differences in yield, and quality should be taken into account when selecting the best approach for processing clinical samples. Hence, when performing comparisons between studies, authors should be aware of the differences among different isolation procedures used. Currently available methods have advantages and disadvantages to consider depending on the approach used as well as any additional stabilizers used such as RNA Later and RNA Protect (Thorn et al., 2005; Schagat et al., 2008; Tavares et al., 2011; Hu et al., 2014; Javadi et al., 2014; Jeffries et al., 2014; Kim et al., 2014).

During these past years, the majority of research efforts of NGS applications were focused on the understanding of the etiology of diseases including complex diseases (Devonshire et al., 2013). Few of the research efforts result in a diagnostic test that is currently being used in a clinical setting. In some occasions, the diagnostic test for a certain disease could be reduced

to a few key genes making it much easier to handle with a simple and cost efficient qPCR or RT-qPCR test (e.g. *BCR-ABL*, *BRCA1*, *BRCA2*, *PKD1*, *PKD2*, retroviruses' RNA). However, in multi-factorial diseases the number of relevant biomarkers is usually too high to be handled by a single PCR assay. For diagnosing complex diseases, many small dedicated assays are being used in combination, such as in AML. Therefore, the analysis of a complete genome, exome or even transcriptome might be considered as a good alternative. Below, we will provide a number of examples showing useful approaches for the analysis of nucleic acids to diagnose inherited or acquired diseases.

Table 7: Some examples of clinical applications for PCR-, microarray- and sequencing-based.

Disease	Tissue	Target genes	Type	Method	Application	Source
Diabetes after kidney transplant	Peripheral blood	ADIPOQ, CCL5	SNPs	qPCR	Research	(Nicoletto et al., 2013)
Hepatitis	Blood plasma	HCV	Retrovirus	qPCR	Commercial kit	www.pcrdiagnostics.eu
Septic Shock	Blood	100 gene signature	Qual. and quant. Expression	µ@	Biomarkers	(Wong et al., 2009)
Hereditary SNRS	Blood	446+ gene panel	SNPs, SNVs	WES	Diagnostics and research	Bristol Genetics, (McCarthy et al., 2013)
Transplant rejection	Peripheral blood	*IL1R2, CXCR4, HLA-A*, +7	Expression	RNAseq	Research	(Chen et al., 2013)
ADPKD	Tissue biopsy	*PKD1, PKD2*	SNV, large deletions	WES	Research and diagnostics	(Rossetti et al., 2012; Tan et al., 2014; Eisenberger et al., 2015)

*SRNS steroid-resistant nephrotic syndrome; ADPKD autosomal dominant polycystic kidney disease; µ@ microarray

1.4.2. Clinical applications, cases and potential examples

1.4.2.1. PCR-based applications

The diagnostic use of PCR or RT-PCR is a very common practice and is frequently applied to detect hundreds of different diseases. The majority of clinical centers and diagnostic facilities offer a wide range catalog of services based on PCR kits to routinely diagnose patients. In this thesis, we will mention only few examples of PCR-based tests that are or could be routinely applied to diagnostics. Although PCR-based testing is not within the scope of our studies, PCR-based diagnostics still deserves a place in this chapter for its importance in the field of clinical diagnostics.

Recent advances in PCR technologies have extended the range of possibilities to use PCR for applied diagnostics. These new PCR platforms, considered as the next generation real-time qPCR, include a wide variety of PCR settings such as micro-fluidic chips, digital-PCR, emulsion digital-PCR also known as digital droplet PCR (dPCR), micro-fluidic chip-based

dPCR, and steel chips. Many of these take advantage of micro and nano-fluidics combined with miniaturized systems to produce an increased number of reactions with a significant reduction of reagent used, sample volume, and costs. This allowed the increase of throughput of these platforms to thousands of reactions per sample making these platforms quite suitable for a rapid screening of relatively large series of biomarkers. For a comprehensive review about PCR-based technologies, advances, and advantages, as well as limitations, please refer to Devonshire *et al.*, (2013).

Infectious diseases represent some of the clearest examples of PCR use for diagnostic purposes, and taking advantage of some PCR tests that are available as commercialized kits. The purpose of these tests is to detect a particular nucleic acid sequence in the analyzed sample such as blood and buccal swaps. These tests have the purpose to be qualitative, providing the answer of presence or absence of a specific DNA or RNA, and in some cases, offer quantitative results. Some of these tests for infectious diseases are based on the detection of the RNA of pathogenic viruses such as the human immunodeficiency virus (HIV) or the hepatitis virus. An example of a commercially available test based on a qPCR kit for detecting HIV may be the AmpliSense HIV-Monitor-FRT kit from PCR Diagnostics. This test uses real-time hybridization-fluorescence to detect HIV type 1 RNA in plasma samples. Similar to this test, other commercially available tests exist for detecting infectious diseases targeting DNA or RNA in plasma samples as well as other tissue sources. Some examples of available tests are listed in Table 7as well as additional relevant test details and references. One of these examples, in the field of renal disorders, is the use of a few qPCR test to identify 3 different variants within the adiponectin and CCL5 gene sequences. These, showed to be associated with the onset of diabetes after renal transplantation (Nicoletto et al., 2013).

1.4.2.2. Microarrays-based applications

The technology of micro-arrays is based on the design of multiple DNA probes that are bound on a solid surface such as a glass slide. Micro-array technologies have been widely used in research for measuring gene expression changes and elucidating the relationship between genotypes and phenotypes. They are quite cost effective for profiling gene expression. Micro-arrays have also been used in clinical diagnostics. Some of the most commonly used examples include the detection of copy number variants using SNP-arrays such as the Cytogenetics Whole-Genome array from Affymetrix or the HumanOmni1-Quad

BeadChip, and HumanCytoSNP-12 DNA Analysis BeadChip from Illumina. In general, micro-arrays can be used for general screenings, gene expression profiling, genotyping and many other applications. However, like in PCR based applications, the use of predefined oligonucleotides (probes) is based on previous knowledge availability. Thus, micro-arrays are used for quantification or detection of known sequences and not for the discovery of new variants, transcripts or other unexpected transcriptomic features (Mortazavi et al., 2008).

Table 8: Technical comparison between PCR, micro-array and RNA-Seq methodologies. Adapted from (Mimura et al., 2014), and complemented with Devonshire et al., (2013) and personal experience and information.

	PCR	**Micro-array**	**DNA/RNA sequencing**
Principle	Hybridization	Hybridization	Sequencing
Resolution	25 bp -100 bp	25-120 bp	Single base
Dependence on available knowledge	High	High	Low
Background noise	High	High	Low
Identification of isoforms	Limited	Limited	High
Differentiate between allelic expressions	Limited	Limited	High
Maximum number of samples per run	384	2	≥96 x 8
Maximum number of targeted regions per sample	1728	30,455*/950,000**	Everything transcribed or amplified

**As of number of CCDS genes from RefSeq Genes annotated and the number of hybridization probes designed for expression arrays.*
***As for the number of probes that can be included/designed for Genotyping SNP Arrays v6.0.*

In order to fully illustrate the limitations of micro-array technology, we should briefly present some basic concepts. Micro-array detection is based on hybridization of sample DNA to nucleic acid probes, bound to the surface of a slide. The probes are oligonucleotides with a usual length of 25 to 120 nucleotides. To further measure the quantity of hybridization to each specific probe, the target sequence (DNA or cDNA) is labeled with fluorescent dyes. Then, after an image is taken and processed, signal intensities can be read and converted to normalized values in order to initiate the data analysis. Due to the nature of micro-array probe design, the capabilities of this method are apparently restricted to known sequences and therefore do not allow detection of target sequences beyond the current knowledge. This factor can be a disadvantage for non-model organisms, but diagnostics of well-characterized organisms, such as humans, is feasible, although it relies on the quality of the available bioinformatics data at the moment the micro-array was designed. Micro-arrays can be used for diagnostic transcriptome analysis. If properly designed, micro-arrays will not only

provide information on gene expression and expressed SNPs, but also detect exon junctions, and fusion genes (Stefano, 2014).

Normalization and processing of micro-array data can involve quite complex bioinformatics methodologies and statistics. This is a consequence of the nature of the data produced by this technology that may become a limitation for someone not acquainted in the area. However, significant efforts were put into developing standardized procedures for micro-array analysis. Some of these procedures as well as suggestions, guidelines, metrics, and thresholds among other information, are publicly available under the micro-array quality control (MAQC) website, together with the publications that helped to reach consensus on these procedures. Reader are referred to the Micro-array Quality Control Project for further details (MicroArray Quality Control (MAQC)).

The human genome has been long studied and annotated, making it easier to use the available information to develop micro-array probes for clinical diagnostics. As an example, Agendia N.V. (www.agendia.com) developed clinical tests for complex diseases such as breast and colon cancers, based on gene expression profiling micro-arrays. The Agendia 'MammaPrint' assay can be used to classify different types of breast cancer and to calculate the recurrence risk. This assay was tested on a cohort of 6694 early breast cancer patients in a phase III trial (MINDACT), to investigate the utility of the MammaPrint 70 gene signature for adjuvant chemotherapy (Viale et al., 2016). Similarly, another test for colon cancer could also classify different cancer types as well as calculate recurrence risk factors with an 18 gene-signature (Salazar et al., 2011).

Genome wide micro-arrays can yield a global view of known polymorphisms or gene expression. Compared to custom designed arrays, genome wide array analyses provide a better opportunity of resolving complex and heterogeneous clinical syndromes. A review by Wong., (2012) of several approaches for sepsis and septic shock, shows the potential and applications of the genome wide microarray analysis (Wong, 2012). One of the outcomes of the studies carried out with septic shock, was the characterization of a 100-gene expression signature. The correlation of clinical phenotype of these pediatric patients with microarray data showed that this expression signature could classify septic shock of three different phenotypes. Being one of these, a severe phenotype with increased illness severity and higher mortality rate (Wong et al., 2009). The proper identification of phenotype-correlated marker genes may also lead researchers to find potential therapeutic targets. Hence, proper

classification to clinical phenotypes of septic shock would allow for the design of more specific and targeted therapies. With the review of many studies of septic shock showing similar results for genes such as *MMP-8*, highly expressed in patients with septic shock, it was possible to demonstrate that inhibition *MMP-8* resulted in significant improvement of patient survival. Hence, genome wide micro-array approaches showed to provide insights in the pathology of complex diseases, help to classify patients in groups with specific characteristics, and allow the discovery of novel therapeutic targets (Wong, 2012).

1.4.2.3. Sequencing-based diagnostics

Advances in DNA sequencing, and in particular the advances of NGS, have significantly improved the quantity and quality of genomic information that can be obtained from clinical samples. The reduced cost of NGS as well as the increase in throughput made whole genome sequencing (WGS), as well as other NGS applications such as whole exome sequencing (WES) or RNA-Seq, a possible and reliable approach for clinical diagnosis. However, there are still some challenges such, as data storage, management, analysis, and interpretation, that have to be considered for the proper use of this technology in clinical applications (Su et al., 2011).

Many different platforms for massive parallel sequencing were developed. The first example, although currently obsolete, is the 454 Genome Sequencer from Roche Applied Sciences. Also outdated is the SOLID platform from Life Technologies. The still current and also most widely used technology is the Solexa 'Sequencing-by-Synthesis' technology that was acquired by Illumina in 2007. The strength of these technologies relies on a very high throughput at the expense of read accuracy and much shorter read length when compared to the well-known Sanger sequencing. However, the possibilities of use and applications of this technology led to significant scientific discoveries and diagnostic applications (Su et al., 2011). Fortunately, some of the trade-offs are being reduced through continuous platform improvements and developments which resulted in more advanced sequencer versions such as the Ion Torrent and Ion Proton from Life Technologies, and the MiSeq, NextSeq and HiSeq from Illumina. In particular, the HiSeq versions have greatly improved in accuracy and read length in the most recent versions as well as in significantly higher throughput (e.g. HiSeq 4000). Meanwhile, the run time has been decreasing, making it suitable for diagnostic use. Advances, and ongoing efforts to improve these platforms even further have made the HiSeq platforms from Illumina the most widely used NGS sequencers.

Depending on the sequencing platform of preference many options are available for library preparations. The library preparation steps include all transformations the nucleic acids of interest may require prior to being completely ready for sequencing on the platform of choice. In general, NGS library preparations for sequencing-based analysis consist of cDNA synthesis (in the case of RNA) and extension of the DNA/cDNA with specific ligated adapters for sequencing. Furthermore, it is quite common that a minimum quantity of input material is required to ensure a minimal quality. In the case of RNA, standard protocols for RNA isolation from body fluids and tissues yield at least 10ng of RNA, which is often sufficient. While for samples containing degraded RNA, such as FFPE, a minimum of 100ng is strongly recommended (Wan et al., 2014). In addition, many adaptations to library preparation protocols are reported in order to cover different aspects of the complexity of DNA or/and RNA processes and regulations such as many types of DNA mutations, methylation and other epigenomic modifications, post-transcriptional modifications, gene expression, isoforms, regulation, splicing, and degradation (Dominissini et al., 2011; Sánchez-Pla et al., 2012; Feng et al., 2013; McGettigan, 2013). For a better overview of available published protocols please refer to available collections of preparation methods such as the sequencing methods review published from Illumina Technology (www.illumina.com/content/dam/illumina-marketing/documents/products/research_reviews/sequencing-methods-review.pdf).

The overwhelming quantity of data produced per sample, requires advanced bioinformatics analysis to address the wide variety of possible questions. There are many tools and software packages available that can analyze these massive datasets, make inferences from the data, and offer biological interpretations. Despite their differences, there are some data analysis steps that are usually shared amongst the different approaches. Common steps include: quality check of the sequencing data, sequence alignment to a reference genome or *de novo* assembly in some other cases, and the assessment of the specific experimental results in order to finally provide useful diagnostic information (Su et al., 2011; Shyr and Liu, 2013; Wan et al., 2014). It is accepted as good practice, to perform several quality checks at the different steps in the process of analyzing clinical samples. Several authors reviewed different quality measures and how to use them during the downstream analysis. A recent review by Li *et al.*, exposed many sequencing quality checks specific for RNA-Seq experiments including checks assessing raw sequence quality, nucleotide composition, presence of rRNA or tRNA, as well as the presence of other contaminant nucleic acids (Li et al., 2015). Although these checks are

specific for RNA sequencing approaches, there are other important and shared steps with DNA-based sequencing approaches of importance, such as the alignment step of the sequenced reads against the reference genome, or transcriptome. The human genome is nowadays quite complete with the latest version 38 released on June 29th 2014 by the Genome Reference Consortium, patch 10 released January 6th 2017 (GRCh38.p10) (Human genome overview - Genome Reference Consortium). During the alignment, the human genome is used as matching reference for the sequenced reads. RNA-Seq data alignments differ substantially from the DNA-Seq alignments. The nature of read sequences in RNA-Seq provides extra levels of complexity due to the fact that RNA molecules are the product of transcription and post-transcriptional processes such as splicing and RNA editing. The splicing removes part of the transcribed sequences - the introns - leaving the exons present in the sequence. After the library preparation and its fragmentation step, which is an optional step and commonly performed by sonication, some of the shorter reads obtained may come from the region where two exons were joined. In this particular situation, the RNA-Seq aligners have to be flexible enough to be able to map part of the reads to one exon and the other part to another exon, spanning an exon junction (Au, 2015). There are many aligners available that can deal with RNA-Seq data, such as Bowtie2, GSNAP, STAR, and SpliceMap among many others. All RNA-seq aligners can deal with both DNA- and RNA-based sequence reads, but this is not true for most DNA aligners that would struggle with the allocation of split reads. Work has been done to review and report available alignment tools to help users through the, sometimes difficult, decision of selecting the best tools for applications in clinical diagnostics (Shyr and Liu, 2013; Wan et al., 2014). In general, all aligners offer the possibility to modify key parameters in order to adapt their algorithms according to the quality of available data and the question of relevance. Once a decent quality alignment is produced, the proper diagnosis is usually within reach.

Particularly for RNA samples, a common approach is to retrieve transcript abundance, as gene counts, for gene expression profiles or differential expression. However, prior to comparing two RNA-Seq datasets, the raw counts should be normalized to account for some differences introduced by handling during the library preparation steps. Due to this inherent variability, normalization of raw counts is required since these are not directly comparable between or within samples (Dillies et al., 2013). There are many normalization methods, some correcting for gene length, GC content, library size (total number of reads), as well as other bias adjustments. For better understanding of the available normalization procedures,

Dillies et al., compared several normalization methods in order to clearly present their application in the context of RNA-Seq data. In summary, the available DESeq and TMM normalization methods showed to be able to maintain the power to detect differential expressed genes while properly controlling the false-positive rate (Dillies et al., 2013). Another way of normalization to deal with extra biases found in cross-platform or inter-laboratory comparisons relies on the inclusion of synthetic spike-in materials. In some cases these external RNA controls developed by the External RNA Controls Consortium (ERCC) became available for the evaluation of cross-platform performance according to GC content, transcript length, and sequencing accuracy (Devonshire et al., 2013). Extended information on RNA-Seq practices as well as some additional recommendations, benchmarking technology comparisons, reproducibility assessments and evaluations of RNA-Seq for clinical applications was also published by the Sequencing Experiment Quality Control consortium (SEQC). The SEQC project is the third phase of the MAQC and it involves 12 countries, 78 organizations and 180 researchers (http://www.fda.gov/ScienceResearch/BioinformaticsTools).

Regardless of the source of sequencing, based on DNA or RNA, the precision of single base information can be easily used to extract the differences of a given sample from the reference genome sequence. The wide range of available bioinformatics tools offers the possibility to answer various biological and diagnostic questions. However, bioinformatics analysis may not be able to overcome some limitations that we can still face with NGS data such as highly repetitive sequences, 3' biases and biased GC content. In general, the small loss of information due to these limitations is of low impact compared with the significant insights that NGS provides. Repetitive sequences in the human genome are well characterized, making it easier to handle problems related to polymorphic copy number variation in these regions. During the alignment steps, reads that map to many locations of the genome (not uniquely mapped) with equal quality are usually filtered out. The enrichment of 3' end sequences of genes, also known as 3' bias, is a side effect of the fast degradation of mRNAs from the 5’end of the transcript, which may be even more prominent when using poly-A enrichment methods during the library preparation. This effect can be widely avoided by using higher quality RNA, which should be possible in a properly designed diagnostic setting. Additionally, 3' biases will not affect the outcome of DNA-based analysis and some RNA-based analysis, such as gene expression measurement, since it is considered that all transcripts exhibit similar degradation and the same library preparation was performed within

a particular well-controlled experiment. The last limitation, regarding some difficulties of sequencing high GC regions, is a problem that usually results from several causes. First, it is known that some polymerases may have increased difficulties to transcribe high GC content sequences. This, coupled with the inherent high repetitive nature of GC or AT enriched regions, makes these regions somehow tricky to analyze with higher levels of confidence. Though, not all high GC-rich sequences are affected at the same level due to differences in GC percentages and other nucleic acid composition, namely AT (Li et al., 2015). High GC regions will not only affect the resolution of SNPs, SNVs, or other genomic variants, but it may also interfere with measurements in gene expression values within GC-rich regions. Hansen *et al.,* worked on an alternative normalization method to acquaint for the GC content as well as gene length of a particular gene using a conditional quartile normalization (Hansen *et al.,* 2012). However, their method did not outperform other less sophisticated normalization methods (Dillies et al., 2013).

Cancer is commonly regarded as an accumulation of genetic alterations such as single nucleotide variants (SNVs), altered DNA methylation patterns, and chromosomal abnormalities. As a consequence of DNA modifications there may be dysfunctional genes leading to over or under-activity and chimeric transcripts or gene fusions. These alterations may disrupt the proper function of the gene, which may become an oncogene, a malfunctioning tumor suppressor, or an incorrect DNA repair gene. The occurrence of one or more of these genetic alterations may affect cellular growth and lead to tumor development. Since the landscape of cancer is complex, sequencing-based approaches (DNA/RNA-Seq) can be very useful for clinical diagnostic applications, offering a wider range of screening possibilities to check for the whole diversity of cancer-related alterations in a single run (Shyr and Liu, 2013). Many studies have been carried out that contributed in the understanding of molecular determinants of tumor cell types. Cancer characterization is remarkably one of the research fields that has dedicated considerable efforts to adopt sequencing approaches, and in particular RNA-Seq, for research purposes and to assess its potential in clinical applications (Su et al., 2011; Shyr and Liu, 2013; Wan et al., 2014; Chaussabel, 2015). Since the accumulation of genetic alterations may be either inherited or somatically acquired, sequencing approaches becomes a strong complementary approach in screening and diagnostic applications.

Apart from complex diseases such as cancer, NGS showed great potential in other fields of clinical diagnostics such as immunology. Recent tendencies in immunological studies showed that large scale genomics and transcriptomics approaches are becoming more popular options for immunological studies. These global approaches can help to mitigate for instance the limitation of cellular heterogeneity in blood samples, even though, the measurement of cell composition in blood samples is not always perfect (Chaussabel, 2015). Despite this limitation, the use of a global approach, such as transcriptomic analysis of blood samples, has shown to provide very nice advantages towards clinical diagnostics as was seen in early research done on autoimmune disorders, cancer, and infectious diseases (Chaussabel, 2015).

Transplant rejection of heart transplants has been assessed by an endo-myocardial biopsy test. But this is an invasive procedure that is characterized by greater risk of morbidity, discomfort for the patient, tissue sampling errors and late detection of rejection. Due to all these limitations, it was necessary to find an alternative to this type of invasive tests to detect heart transplant rejection, and to more accurately and early adapt patient treatment to avoid transplant rejections. With this in mind, Chen *et al.,* (2013) used a NGS approach to analyze peripheral blood gene expression profiles, monitor the immune system and potentially avoid heart transplant rejection by early detection. For this study 12 patients were analyzed from grade 0 (6 quiescent patients) to grade 2R and 3R (6 rejection patients). The results were validated by qPCR of 47 individuals from three different rejection groups. A total of 10 genes (Table 7) were identified which provided a signature of high risk for severe rejection. This 10-gene signature was also tested to be effective in other organ transplants (Chen et al., 2013).

In the case of patients with hereditary steroid-resistant nephrotic syndrome (SRNS), McCarthy et al., (2013) showed the potential of NGS to investigate the genetic variations of this group of patients in a single step. Causa variants were identified using a WES panel of 446 genes which included the screening of the 24 genes currently associated with SRNS. The screening of >350 patients with this approach, showed to be reliable to identify previously known disease-associated variants in genes such as *NPHS1*, *NPHS2*, and *PLCe1* as well as the potential of identifying *de novo* variants in unexpected genes such as *COQ2* and *COL4A4* (McCarthy et al., 2013).

Research work in the field of renal diseases also showed the potential of NGS to identify gene expression profiles, gene pathways, and alternative splicing linked to *TGFB* and *SMAD3*

signaling in chronic kidney diseases (CKD) (Zhou et al., 2015). In summary, research findings using animal models provided insight information about *SMAD3* signaling and its function in renal injury, as well as highlighting potential targets for CKD therapies (Zhou et al., 2015). The study of *SMAD3*-dependent renal injury was performed using total RNA of kidneys from mice models with *SMAD3* wild type and knockouts for immune and non-immune mediated CKD (anti-glomerular basement membrane glomerulonephritis, and obstructive nephropathy, respectively). Zhou *et al*., reported 9 differentially expressed genes linked to *SMAD3* (*IGHG1*, *IGHG2C*, *IGKV12-41*, *IGHV14-3*, *IGHV5-6*, *IGHG2B*, *UGT2B37*, *SLC22A19* and *MFSD2A*) and showed that renal injury transcriptomes may mediate pathogenesis of CKD.

Global as well as targeted NGS approaches showed to be suitable for unravelling the insights and progression of complex diseases such as CKD (Zhou et al., 2015). In addition, NGS approaches can facilitate the identification of inherited pathogenic mutations in key disease-associated genes such as *PKD1* or *PKD2* for ADPKD (Rossetti et al., 2012). Overall, NGS approaches offer an alternative to classic/current diagnostics that deserve to be part of the daily basis of applied and personalized medicine and can be beneficial in the field of renal disorders.

1.5. Scope

The general aim of this thesis is to bridge between two very well stablished and distinct specializations, nephrology and sequencing data analysis, with the objective of providing new and advanced tools for the better understanding of renal diseases. With this purpose, the studies described within this thesis will use different cutting edge sequencing approaches to exemplify the use of NGS in three distinct renal disorders: ADPKD, AKI, and glomerulosclerosis. We will discuss the potential advantages of using sequencing approaches in each different case.

ADPKD: In Chapter 2 we provide a new methodology for the diagnosis of ADPKD based on cutting edge NGS approaches with long-read single molecule sequencing (PacBio). Particularly, we show that long-read sequencing can overcome the difficulties and complexities associated with *PKD1*.

AKI: In Chapter 3 we provide a clear example of the use of RNA-sequencing for expression profiling of a complex disease such as AKI. In addition, we show the potential uses of

genome wide transcriptomics approaches to elucidate a complex genetic mechanism, which includes signalling pathways that are activated in the presence or absence of a given gene such as *Gdf15*. In return, we show that AKI expression profiling with NGS approaches can identify the driving transcription factors of an AKI phenotype.

Glomerulosclerosis: In Chapter 4, we evaluate RNA-sequencing of FFPE tissue with the aim of characterizing glomerular expression profiles of sclerotic glomeruli isolated with LCM. In addition, we define the minimal requirements for a successful sequencing approach of RNA isolated from FFPE samples.

Finally, we will elaborate on the future perspectives of sequencing technologies and their potential uses in future diagnostics applications.

1.6. References

Antonarakis SE, Beckmann JS. 2006. Mendelian disorders deserve more attention. Nat Rev Genet 7:277–282.

Au KF. 2015. Accurate Mapping of RNA-Seq Data. In: Picardi E, editor. RNA Bioinformatics, New York, NY: Springer New York, p 147–161.

Barua M, Cil O, Paterson AD, Wang K, He N, Dicks E, Parfrey P, Pei Y. 2009. Family History of Renal Disease Severity Predicts the Mutated Gene in ADPKD. J Am Soc Nephrol 20:1833–1838.

Bellomo R, Kellum JA, Ronco C. 2012. Acute kidney injury. Lancet Lond Engl 380:756–766.

Bozic I, Gerold JM, Nowak MA. 2016. Quantifying Clonal and Subclonal Passenger Mutations in Cancer Evolution. PLOS Comput Biol 12:e1004731.

Breit SN, Carrero JJ, Tsai VW-W, Yagoutifam N, Luo W, Kuffner T, Bauskin AR, Wu L, Jiang L, Barany P, Heimburger O, Murikami M-A, et al. 2012. Macrophage inhibitory cytokine-1 (MIC-1/GDF15) and mortality in end-stage renal disease. Nephrol Dial Transplant Off Publ Eur Dial Transpl Assoc - Eur Ren Assoc 27:70–75.

Brinkman RR, Dubé M-P, Rouleau GA, Orr AC, Samuels ME. 2006. Human monogenic disorders — a source of novel drug targets. Nat Rev Genet 7:249–260.

Castelli M, De Pascalis C, Distefano G, Ducano N, Oldani A, Lanzetti L, Boletta A. 2015. Regulation of the microtubular cytoskeleton by Polycystin-1 favors focal adhesions turnover to modulate cell adhesion and migration. BMC Cell Biol 16:15.

Chaisson MJP, Huddleston J, Dennis MY, Sudmant PH, Malig M, Hormozdiari F, Antonacci F, Surti U, Sandstrom R, Boitano M, Landolin JM, Stamatoyannopoulos JA, et al. 2015. Resolving the complexity of the human genome using single-molecule sequencing. Nature 517:608–611.

Chaussabel D. 2015. Assessment of immune status using blood transcriptomics and potential implications for global health. Semin Immunol 27:58–66.

Chen Y, Zhang H, Xiao X, Jia Y, Wu W, Liu L, Jiang J, Zhu B, Meng X, Chen W. 2013. Peripheral blood transcriptome sequencing reveals rejection-relevant genes in long-term heart transplantation. Int J Cardiol 168:2726–2733.

Del Prete MJ, Vernal R, Dolznig H, Mullner EW, Garcia-Sanz JA. 2007. Isolation of polysome-bound mRNA from solid tissues amenable for RT-PCR and profiling experiments. RNA 13:414–421.

Denic A, Lieske JC, Chakkera HA, Poggio ED, Alexander MP, Singh P, Kremers WK, Lerman LO, Rule AD. 2017. The Substantial Loss of Nephrons in Healthy Human Kidneys with Aging. J Am Soc Nephrol JASN 28:313–320.

Devonshire AS, Sanders R, Wilkes TM, Taylor MS, Foy CA, Huggett JF. 2013. Application of next generation qPCR and sequencing platforms to mRNA biomarker analysis. Methods 59:89–100.

Dillies M-A, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, Keime C, Marot G, Castel D, Estelle J, Guernec G, Jagla B, et al. 2013. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Brief Bioinform 14:671–683.

Ding M, Bullotta A, Caruso L, Gupta P, Rinaldo CR, Chen Y. 2011. An optimized sensitive method for quantitation of DNA/RNA viruses in heparinized and cyropreserved plasma. J Virol Methods 176:1–8.

Doi K, Okamoto K, Negishi K, Suzuki Y, Nakao A, Fujita T, Toda A, Yokomizo T, Kita Y, Kihara Y, Ishii S, Shimizu T, et al. 2006. Attenuation of Folic Acid-Induced Renal Inflammatory Injury in Platelet-Activating Factor Receptor-Deficient Mice. Am J Pathol 168:1413–1424.

Dominissini D, Moshitch-Moshkovitz S, Amariglio N, Rechavi G. 2011. Adenosine-to-inosine RNA editing meets cancer. Carcinogenesis 32:1569–1577.

Eisenberger T, Decker C, Hiersche M, Hamann RC, Decker E, Neuber S, Frank V, Bolz HJ, Fehrenbach H, Pape L, Toenshoff B, Mache C, et al. 2015. An Efficient and Comprehensive Strategy for Genetic Diagnostics of Polycystic Kidney Disease. PLOS ONE 10:e0116680.

Fang T-C, Alison MR, Cook HT, Jeffery R, Wright NA, Poulsom R. 2005. Proliferation of bone marrow-derived cells contributes to regeneration after folic acid-induced acute tubular injury. J Am Soc Nephrol JASN 16:1723–1732.

Feng H, Qin Z, Zhang X. 2013. Opportunities and methods for studying alternative splicing in cancer with RNA-Seq. Cancer Lett 340:179–191.

Gainullin VG, Hopp K, Ward CJ, Hommerding CJ, Harris PC. 2015. Polycystin-1 maturation requires polycystin-2 in a dose-dependent manner. J Clin Invest 125:607–620.

Gansevoort RT, Arici M, Benzing T, Birn H, Capasso G, Covic A, Devuyst O, Drechsler C, Eckardt K-U, Emma F, Knebelmann B, Le Meur Y, et al. 2016. Recommendations for the use of tolvaptan in autosomal dominant polycystic kidney disease: a position statement on behalf of the ERA-EDTA Working Groups on Inherited Kidney Disorders and European Renal Best Practice. Nephrol Dial Transplant 31:337–348.

Go AS, Chertow GM, Fan D, McCulloch CE, Hsu C. 2004. Chronic Kidney Disease and the Risks of Death, Cardiovascular Events, and Hospitalization. N Engl J Med 351:1296–1305.

Griffionen M, Arindrarto W, Borràs DM, Diessen SAME van, Meijden ED van der, Honders MW, Alloul M, Jedema I, Kroes WGM, Valk PJM, Janssen B, Bergen CAM van, et al. 2016. Whole Transcriptome Sequencing (RNAseq) As a Comprehensive, Cost-Efficient Diagnostic Tool for Acute Myeloid Leukemia. In: 58th Anual Meeting and Exposition of the Aerican Society of Hematology, San Diego.

Guo X, Zheng S, Dang H, Pace RG, Stonebraker JR, Jones CD, Boellmann F, Yuan G, Haridass P, Fedrigo O, Corcoran DL, Seibold MA, et al. 2013. Genome Reference and Sequence Variation in the Large Repetitive Central Exon of Human MUC5AC. Am J Respir Cell Mol Biol 50:223–232.

Gutierrez-Arcelus M, Ongen H, Lappalainen T, Montgomery SB, Buil A, Yurovsky A, Bryois J, Padioleau I, Romano L, Planchon A, others. 2015. Tissue-Specific Effects of Genetic and Epigenetic Variation on Gene Regulation and Splicing. PLoS Genet 11:e1004958.

Hall JE, Guyton AC. 2011. Guyton and Hall Textbook of Medical Physiology. Philadelphia, Pa: Saunders/Elsevier. 1091 p.

Han KH, Kim SH. 2016. Recent Advances in Treatments of Primary Focal Segmental Glomerulosclerosis in Children. BioMed Res Int 2016:.

Harris PC, Rossetti S. 2010. Molecular diagnostics for autosomal dominant polycystic kidney disease. Nat Rev Nephrol 6:197–206.

Harris PC, Torres VE. 1993. Polycystic Kidney Disease, Autosomal Dominant. In: Pagon RA, Adam MP, Ardinger HH, Wallace SE, Amemiya A, Bean LJ, Bird TD, Ledbetter N, Mefford HC, Smith RJ, Stephens K, editors. GeneReviews(®), Seattle (WA): University of Washington, Seattle,.

Harris PC, Torres VE. 2014. Genetic mechanisms and signaling pathways in autosomal dominant polycystic kidney disease. J Clin Invest 124:2315–2324.

Harty J. 2014. Prevention and Management of Acute Kidney Injury. Ulster Med J 83:149–157.

Herder C, Karakas M, Koenig W. 2011. Biomarkers for the Prediction of Type 2 Diabetes and Cardiovascular Disease. Clin Pharmacol Ther 90:52–66.

Hill NR, Fatoba ST, Oke JL, Hirst JA, O'Callaghan CA, Lasserson DS, Hobbs FDR. 2016. Global Prevalence of Chronic Kidney Disease – A Systematic Review and Meta-Analysis. PLoS ONE 11:.

Hoy WE, Ingelfinger JR, Hallan S, Hughson MD, Mott SA, Bertram JF. 2010. The early development of the kidney and implications for future health. J Dev Orig Health Dis 1:216–233.

Hu Y-C, Zhang Q, Huang Y-H, Liu Y-F, Chen H-L. 2014. Comparison of two methods to extract DNA from formalin-fixed, paraffin-embedded tissues and their impact on EGFR mutation detection in non-small cell lung carcinoma. Asian Pac J Cancer Prev APJCP 15:2733–2737.

Human genome overview - Genome Reference Consortium.

Javadi A, Shamaei M, Mohammadi Ziazi L, Pourabdollah M, Dorudinia A, Seyedmehdi SM, Karimi S. 2014. Qualification Study of Two Genomic DNA Extraction Methods in Different Clinical Samples. Tanaffos 13:41–47.

Jeffries MKS, Kiss AJ, Smith AW, Oris JT. 2014. A comparison of commercially-available automated and manual extraction kits for the isolation of total RNA from small tissue samples. BMC Biotechnol 14:94.

Kiffel J, Rahimzada Y, Trachtman H. 2011. FOCAL SEGMENTAL GLOMERULOSCLEROSIS AND CHRONIC KIDNEY DISASE IN PEDIATRIC PATIENTS. Adv Chronic Kidney Dis 18:332–338.

Kim J-H, Jin H-O, Park J-A, Chang YH, Hong YJ, Lee JK. 2014. Comparison of three different kits for extraction of high-quality RNA from frozen blood. SpringerPlus 3:76.

Koeppen BM, Stanton BA. 2013. Renal Physiology. Philadelphia, PA: Elsevier Mosby. 240 p.

Laver TW, Caswell RC, Moore KA, Poschmann J, Johnson MB, Owens MM, Ellard S, Paszkiewicz KH, Weedon MN. 2016. Pitfalls of haplotype phasing from amplicon-based long-read sequencing. Sci Rep 6:21746.

Lee H, Schatz MC. 2012. Genomic dark matter: the reliability of short read mapping illustrated by the genome mappability score. Bioinformatics 28:2097–2105.

Leonhard WN, Wal A van der, Novalic Z, Kunnen SJ, Gansevoort RT, Breuning MH, Heer E de, Peters DJM. 2011. Curcumin inhibits cystogenesis by simultaneous interference of multiple signaling pathways: in vivo evidence from a Pkd1-deletion model. Am J Physiol Renal Physiol 300:F1193-1202.

Li X, Nair A, Wang S, Wang L. 2015. Quality control of RNA-seq experiments. Methods Mol Biol Clifton NJ 1269:137–146.

Loomis EW, Eid JS, Peluso P, Yin J, Hickey L, Rank D, McCalmon S, Hagerman RJ, Tassone F, Hagerman PJ. 2013. Sequencing the unsequenceable: Expanded CGG-repeat alleles of the fragile X gene. Genome Res 23:121–128.

Mallawaarachchi AC, Hort Y, Cowley MJ, McCabe MJ, Minoche A, Dinger ME, Shine J, Furlong TJ. 2016. Whole-genome sequencing overcomes pseudogene homology to diagnose autosomal dominant polycystic kidney disease. Eur J Hum Genet 24:1584–1590.

Mardis ER. 2013. Next-Generation Sequencing Platforms. Annu Rev Anal Chem 6:287–303.

Mazagova M, Buikema H, Buiten A van, Duin M, Goris M, Sandovici M, Henning RH, Deelman LE. 2013. Genetic deletion of growth differentiation factor 15 augments renal damage in both type 1 and type 2 models of diabetes. Am J Physiol Renal Physiol 305:F1249-1264.

McCarthy HJ, Bierzynska A, Wherlock M, Ognjanovic M, Kerecuk L, Hegde S, Feather S, Gilbert RD, Krischock L, Jones C, Sinha MD, Webb NJA, et al. 2013. Simultaneous Sequencing of 24 Genes Associated with Steroid-Resistant Nephrotic Syndrome. Clin J Am Soc Nephrol 8:637–648.

McGettigan PA. 2013. Transcriptomics in the RNA-seq era. Curr Opin Chem Biol 17:4–11.

Metz-Kurschel U, Kurschel E, Wagner K, Aulbert E, Graben N, Philipp T. 1990. Folate nephropathy occurring during cytotoxic chemotherapy with high-dose folinic acid and 5-fluorouracil. Ren Fail 12:93–97.

MicroArray Quality Control (MAQC).

Mimura I, Kanki Y, Kodama T, Nangaku M. 2014. Revolution of nephrology research by deep sequencing: ChIP-seq and RNA-seq. Kidney Int 85:31–38.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5:621–628.

Nicoletto BB, Souza GC, Fonseca NKO, Centenaro A, Manfro RC, Canani LHS, Gonçalves LFS. 2013. Association between 276G/T adiponectin gene polymorphism and new-onset diabetes after kidney transplantation. Transplantation 96:1059–1064.

O'Connor TP, Crystal RG. 2006. Genetic medicines: treatment strategies for hereditary disorders. Nat Rev Genet 7:261–276.

Oliver GR, Hart SN, Klee EW. 2015. Bioinformatics for Clinical Next Generation Sequencing. Clin Chem 61:124–135.

Ortega A, Rámila D, Ardura JA, Esteban V, Ruiz-Ortega M, Barat A, Gazapo R, Bosch RJ, Esbrit P. 2006. Role of parathyroid hormone-related protein in tubulointerstitial apoptosis and fibrosis after folic acid-induced nephrotoxicity. J Am Soc Nephrol JASN 17:1594–1603.

Petrov A, Wu T, Puglisi EV, Puglisi JD. 2013. RNA Purification by Preparative Polyacrylamide Gel Electrophoresis. Methods in Enzymology, Elsevier, p 315–330.

Qiao W, Yang Y, Sebra R, Mendiratta G, Gaedigk A, Desnick RJ, Scott SA. 2016. Long-Read Single Molecule Real-Time Full Gene Sequencing of Cytochrome P450-2D6: Human Mutation. Hum Mutat 37:315–323.

Ross JS, Hatzis C, Symmans WF, Pusztai L, Hortobagyi GN. 2008. Commercialized multigene predictors of clinical outcome for breast cancer. The Oncologist 13:477–493.

Rossetti S, Consugar MB, Chapman AB, Torres VE, Guay-Woodford LM, Grantham JJ, Bennett WM, Meyers CM, Walker DL, Bae K, Zhang Q, Thompson PA, et al. 2007. Comprehensive Molecular Diagnostics in Autosomal Dominant Polycystic Kidney Disease. J Am Soc Nephrol 18:2143–2160.

Rossetti S, Hopp K, Sikkink RA, Sundsbak JL, Lee YK, Kubly V, Eckloff BW, Ward CJ, Winearls CG, Torres VE, Harris PC. 2012. Identification of Gene Mutations in Autosomal Dominant Polycystic Kidney Disease through Targeted Resequencing. J Am Soc Nephrol JASN 23:915–933.

Salazar R, Roepman P, Capella G, Moreno V, Simon I, Dreezen C, Lopez-Doriga A, Santos C, Marijnen C, Westerga J, Bruin S, Kerr D, et al. 2011. Gene Expression Signature to Improve Prognosis Prediction of Stage II and III Colorectal Cancer. J Clin Oncol 29:17–24.

Sánchez-Pla A, Reverter F, Ruíz de Villa MC, Comabella M. 2012. Transcriptomics: mRNA and alternative splicing. J Neuroimmunol 248:23–31.

Sasagawa Y, Nikaido I, Hayashi T, Danno H, Uno KD, Imai T, Ueda HR. 2013. Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity. Genome Biol 14:R31.

Schagat T, Kiak L, Mandrekar M. 2008.

Sequencing Methods Review.

Shyr D, Liu Q. 2013. Next generation sequencing in cancer research and clinical application. Biol Proced Online 15:.

Spithoven EM, Kramer A, Meijer E, Orskov B, Wanner C, Abad JM, Areste N, Alonso de la Torre R, Caskey F, Couchoud C, Finne P, Heaf J, et al. 2014. Renal replacement therapy

for autosomal dominant polycystic kidney disease (ADPKD) in Europe: prevalence and survival--an analysis of data from the ERA-EDTA Registry. Nephrol Dial Transplant 29:iv15-iv25.

Stefano GB. 2014. Comparing Bioinformatic Gene Expression Profiling Methods: Microarray and RNA-Seq. Med Sci Monit Basic Res 20:138–142.

Steinberg KM, Schneider VA, Graves-Lindsay TA, Fulton RS, Agarwala R, Huddleston J, Shiryev SA, Morgulis A, Surti U, Warren WC, Church DM, Eichler EE, et al. 2014. Single haplotype assembly of the human genome from a hydatidiform mole. Genome Res 24:2066–2076.

Su Z, Ning B, Fang H, Hong H, Perkins R, Tong W, Shi L. 2011. Next-generation sequencing and its applications in molecular diagnostics.

Tan AY, Michaeel A, Liu G, Elemento O, Blumenfeld J, Donahue S, Parker T, Levine D, Rennert H. 2014. Molecular diagnosis of autosomal dominant polycystic kidney disease using next-generation sequencing. J Mol Diagn 16:216–228.

Taussig DC, Vargaftig J, Miraki-Moud F, Griessinger E, Sharrock K, Luke T, Lillington D, Oakervee H, Cavenagh J, Agrawal SG, Lister TA, Gribben JG, et al. 2010. Leukemia-initiating cells from some acute myeloid leukemia patients with mutated nucleophosmin reside in the CD34- fraction. Blood 115:1976–1984.

Tavares L, Alves PM, Ferreira RB, Santos CN. 2011. Comparison of different methods for DNA-free RNA isolation from SK-N-MC neuroblastoma. BMC Res Notes 4:3.

Thorn I, Olsson-Stromberg U, Ohlsen C, Jonsson A-M, Klangby U, Simonsson B, Barbany G. 2005. The impact of RNA stabilization on minimal residual disease assessment in chronic myeloid leukemia. haematologica 90:1471–1476.

Todd R, Kuo MWLWP. 2002. Gene expression profiling using laser capture microdissection. Expert Rev Mol Diagn 2:497–507.

Trujillano D, Bullich G, Ossowski S, Ballarín J, Torra R, Estivill X, Ars E. 2014. Diagnosis of autosomal dominant polycystic kidney disease using efficient PKD1 and PKD2 targeted next-generation sequencing. Mol Genet Genomic Med 2:412–421.

Vallon V. 2016. Tubular Transport in Acute Kidney Injury: Relevance for Diagnosis, Prognosis and Intervention. Nephron 134:160–166.

Viale G, Slaets L, Snoo FA de, Bogaerts J, Russo L, Veer L van't, Rutgers EJT, Piccart-Gebhart MJ, Stork-Sloots L, Dell'Orto P, Glas AM, Cardoso F. 2016. Discordant assessment of tumor biomarkers by histopathological and molecular assays in the EORTC randomized controlled 10041/BIG 03-04 MINDACT trial breast cancer: Intratumoral heterogeneity and DCIS or normal tissue components are unlikely to be the cause of discordance. Breast Cancer Res Treat 155:463–469.

Walker SE, Lorsch J. 2013. RNA Purification – Precipitation Methods. Methods in Enzymology, Elsevier, p 337–343.

Wan M, Wang J, Gao X, Sklar J. 2014. RNA Sequencing and its Applications in Cancer Diagnosis and Targeted Therapy. North Am J Med Sci 7:.

Wetmore JB, Collins AJ. 2016. Global challenges posed by the growth of end-stage renal disease. Ren Replace Ther 2:15.

Wong HR. 2012. Clinical review: sepsis and septic shock–the potential of gene arrays. Crit Care 16:204.

Wong HR, Cvijanovich N, Lin R, Allen GL, Thomas NJ, Willson DF, Freishtat RJ, Anas N, Meyer K, Checchia PA, Monaco M, Odom K, et al. 2009. Identification of pediatric septic shock subclasses based on genome-wide expression profiling. BMC Med 7:34.

Yang L, Besschetnova TY, Brooks CR, Shah JV, Bonventre JV. 2010. Epithelial cell cycle arrest in G2/M mediates kidney fibrosis after injury. Nat Med 16:535–143.

Yu ASL, El-Ters M, Winklhofer FT. 2015. Clinical Trials in Autosomal Dominant Polycystic Kidney Disease. In: Li X, editor. Polycystic Kidney Disease, Brisbane (AU): Codon Publications,.

Zhou Q, Xiong Y, Huang XR, Tang P, Yu X, Lan HY. 2015. Identification of Genes Associated with Smad3-dependent Renal Injury by RNA-seq-based Transcriptome Analysis. Sci Rep 5:17901.

Chapter 2

2. Detecting PKD1 variants in Polycystic Kidney Disease patients by single-molecule long-read sequencing

Daniel M. Borràs[1,2,3], Rolf Vossen[4], Michael Liem[4], Henk P.J. Buermans[4], Hans Dauwerse[5], Dave van Heusden[5], Ron T. Gansevoort[6,#], Johan T. den Dunnen[4,5,7], Bart Janssen[1], Dorien J.M. Peters[5,#], Monique Losekoot[7,#], Seyed Yahya Anvar[4,5,]*

[1]*GenomeScan B.V, Plesmanlaan 1d, 2333BZ Leiden, The Netherlands.*

[2]*Institut National de la Santé et de la Recherche Médicale (INSERM), Institut of Cardiovascular and Metabolic Disease, Toulouse, France.*

[3]*Université Toulouse III Paul-Sabatier, Toulouse, France.*

[4]*Leiden Genome Technology Center (LGTC), Department of Human Genetics, Leiden University Medical Center (LUMC), Leiden, The Netherlands.*

[5]*Department of Human Genetics, Leiden University Medical Center (LUMC), Leiden, The Netherlands.*

[6]*Department of Nephrology, University Hospital Groningen, University Medical Center Groningen, PO Box 30.001, 9700 RB, Groningen, The Netherlands.*

[7]*Department of Clinical Genetics, Leiden University Medical Center (LUMC), Leiden, The Netherlands.*

[#]*On behalf of the DIPAK consortium.*

[*]*Corresponding author: SY Anvar; E-mail: s.y.anvar@lumc.nl*

Grant Sponsor: *The research leading to these results has received funding from the European Union's Seventh Framework Programme FP7/2007-2013 under grant agreement FP7-PEOPLE-2013-ITN-608332.*

2.1. Abstract

A genetic diagnosis of autosomal dominant polycystic kidney disease (ADPKD) is challenging due to allelic heterogeneity, high GC-content, and homology of the *PKD1* gene with six pseudogenes. Short-read next-generation sequencing (NGS) approaches, such as whole genome (WGS) and whole exome sequencing (WES), often fail at reliably characterizing complex regions such as *PKD1*. However, long-read single-molecule sequencing has been shown to be an alternative strategy that could overcome *PKD1* complexities and discriminate between homologous regions of *PKD1* and its pseudogenes. In this study, we present the increased power of resolution for complex regions using long-read sequencing to characterize a cohort of 19 patients with ADPKD. Our approach provided high sensitivity in identifying *PKD1* pathogenic variants, diagnosing 94.7% of the patients. We show that reliable screening of ADPKD patients in a single test without interference of *PKD1* homologous sequences, commonly introduced by residual amplification of *PKD1* pseudogenes, by direct long-read sequencing is now possible. This strategy can be implemented in diagnostics and is highly suitable to sequence and resolve complex genomic regions that are of clinical relevance.

2.2. Keywords

DNA diagnostics, *PKD1*, ADPKD, complex genomic regions, variant detection, Single-Molecule Real-Time sequencing, long-read sequencing, PacBio.

2.3. Introduction

DNA sequencing technologies have been widely applied in biomedical and biological research as well as diagnostics. Relatively low cost and high-throughput are major advantages of next-generation sequencing (NGS) over standard diagnostic assays (Su et al., 2011; Mardis, 2013; Oliver et al., 2015). However, despite widespread use of NGS-based diagnostics strategies (Ozsolak and Milos, 2011; de Ligt et al., 2012; Chang and Li, 2013; von and Huber, 2013; Yang et al., 2013; Dewey et al., 2014; LaDuca et al., 2014; Codina-Solà et al., 2015; Sun et al., 2015; Willig et al., 2015), short-read sequencing approaches such as whole genome (WGS) and whole exome sequencing (WES), often fail at reliably characterizing complex regions of the human genome (Lee and Schatz, 2012; Chaisson et al., 2015). These regions are often associated with extreme GC-content, segmental duplications (SDs), low complexity sequences and gaps in the human reference sequence (Lee and Schatz, 2012; Steinberg et al., 2014; Chaisson et al., 2015).

Single-molecule long-read sequencing can improve our understanding of genetic variations in complex but clinically relevant genomic regions (Guo et al., 2013; Loomis et al., 2013; Laver et al., 2016; Qiao et al., 2016).

In this study, we aim to show the value of single-molecule long-read sequencing as a tool to characterize genetic variants associated with autosomal dominant polycystic kidney disease (ADPKD). ADPKD is a common inherited disease that accounts for 5% to 10% of end-stage renal disease (ESRD) (Harris and Rossetti, 2010; Spithoven et al., 2014). Most ADPKD pathogenic variants occur in *PKD1* (MIM* 601313) and *PKD2* (MIM* 173910) genes with a reported prevalence of 85% and 15%, respectively (Barua et al., 2009; Harris and Rossetti, 2010). The mutation spectrums in *PKD1* and *PKD2* are highly heterogeneous, with no mutation hotspots present, indicating that pathogenic variants in either *PKD1* or *PKD2* are usually private (Gout et al., 2007; Harris and Rossetti, 2010). The screening of *PKD1* is challenging due to difficulties in amplification and low resolution of its complex locus (Rossetti et al., 2007; Tan et al., 2009; Qi et al., 2013). This is partly due to its high homology for most of *PKD1* sequence with six pseudogenes as well as high GC content (Rossetti et al., 2007; Tan et al., 2009; Qi et al., 2013). In this study, we used *PKD1* as an excellent example of a challenging and complex locus.

Several attempts have been made to improve the screening of *PKD1* gene by using short-read NGS approaches to replace the standard diagnostics based on Sanger sequencing and multiplex ligation-dependent probe amplification (MLPA) assays (Rossetti et al., 2012; Qi et al., 2013; Tan et al., 2014; Trujillano et al., 2014; Eisenberger et al., 2015; Mallawaarachchi et al., 2016). These strategies provided a clear diagnosis with high sensitivity and specificity (97% - 100%) for 115 out of 183 (Rossetti et al., 2012), 16 out of 25 (Tan et al., 2014), 10 out of 12 (Trujillano et al., 2014), 35 out of 55 (Eisenberger et al., 2015), and 24 out of 28 (Mallawaarachchi et al., 2016) screened ADPKD patients. Duplicated and high GC-content genomic regions, such as that of *PKD1* gene, can lead to ambiguous identification of variants when analyzed with short-read NGS strategies (Lee and Schatz, 2012). These ambiguities produced low true positive variant detection rates of 28% to 50% for the duplicated region of *PKD1* (Qi et al., 2013), and many false positives (10%) due to misalignments, low quality alignments and contamination by residual amplification of pseudogenes (Rossetti et al., 2012). Hence, diagnostic assays based on NGS short-reads (e.g. Sanger or Illumina) may not be fully suited for reliable ADPKD diagnostics.

In this study, we utilized the single-molecule long-read Pacific Biosciences RSII (PacBio) sequencing technology to assess its potential value in molecular diagnostics of ADPKD patients. We show that direct sequencing of long-range PCR (LR-PCR) products eliminates the interference of residual amplification of *PKD1* pseudogenes, as well as alignment ambiguities. This also enabled a reliable identification of pathogenic variants, from single nucleotide variants (SNVs) to large deletions.

2.4. Materials and Methods

2.4.1. Selection of Subjects and DNA Isolation

Nineteen genotyped patient samples were selected for this study from the diagnostic laboratory in which at least one pathogenic mutation was detected by Sanger sequencing or MLPA. The selection aimed to include different types of mutations (eg. SNVs, as well as small and larger insertions and deletions (indels)) that are located in exons or in immediately flanking intronic sequences, for both the duplicated regions as well as the unique part of *PKD1*. Although *PKD2* is not a complex gene and is not the focus of this study, the sequencing of LR-PCR fragments for *PKD2* was performed as a proof of principle of long-read sequencing and detection of variants also for *PKD2*. Genomic DNA isolation was performed from peripheral blood samples using PUREGENE^{TM} nucleic acid purification chemistry on the AUTOPURE LS 98 Instrument (Qiagen).

2.4.2. Long-read sequencing and variant identification for ADPKD genes

2.4.2.1. LR-PCR amplification

To cover the entire *PKD1* and *PKD2* coding regions (including exon boundaries), a total of 5 and 9 LR-PCR fragments were designed, respectively. Primers were optimized to produce amplicons of similar sizes (>4Kb) that could be pooled to improve sequencing efficiency and loading capacity for SMRT sequencing (Supp. Table S1; Supp. Figure S1). The major part of *PKD1* intron 1 was excluded from the design due to its large size and the lack of previously reported pathogenic variants in this region. Fragments were amplified from 50ng of genomic DNA using 1x Extensor^{TM} Hi-Fidelity Long Range PCR Master Mix (Thermo Scientific) on a 25 µl of PCR reaction volume with 200 nM of M13-tagged primers. Initial denaturation was performed for 10 min at 98°, followed by 35 cycles of 15 secs at 98° and 10 min at 68°. Final extension was 10 min at 68°. Products were size selected using the BluePippin DNA size selection system to

classify them in 3 different groups of sizes 4.3-6-1 Kb, 7.1-7.5 Kb, and 7.6-8.1 KB (Supp. Table S1; Supp. Figure S1). Fragments of equal size were pooled equimolar, and were visually inspected by band intensity on agarose gel. Finally, all pools were purified with a 0.6x v/v ratio of AMPure XP Beads (Beckman-Coulter).

2.4.2.2. SMRT-sequencing library preparation

Sample indexes for patient tracking were added to the LR-PCR fragments using an additional five cycle PCR with the previous LR-PCR conditions. Barcoded pools were then purified with AMPure XP Beads, and pooled equimolar according to their size. Molar concentration was verified on a Bioanalyzer 12000 chip (Agilent). For each barcoded pool, a SMRT-bell library

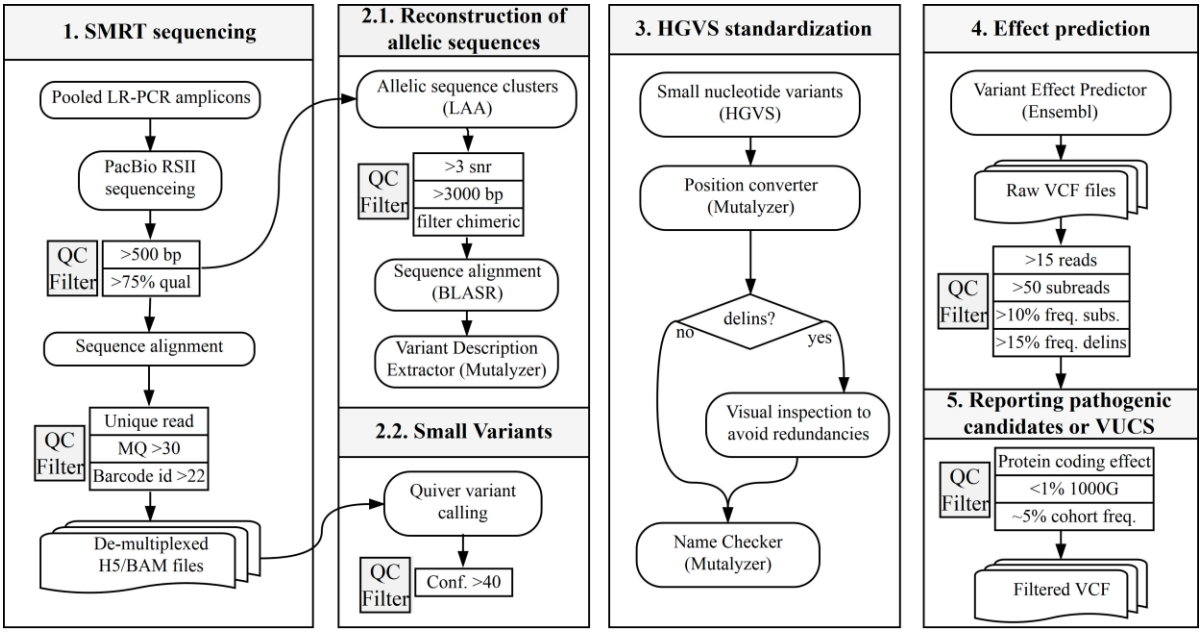


Figure 1: Flowchart of the applied analytical approach for the identification of potentially pathogenic variants and VUCS in ADPKD patient samples. Key processes in the workflow describe details and thresholds used for (*1.*) sequencing of pooled LR-PCR amplified fragments with PacBio RSII and postprocessing of reads including alignments and read quality filters; (*2.*) identification of variants using two independent strategies including the reconstruction of allelic sequences, and small variant calling using Quiver; (*3.*) standardization of variant nomenclature to represent a correct HGVS description and facilitate the comparison between datasets; (*4.*) enrichment of variant annotations with VEP (including effect prediction, ClinVar, SIFT, PolyPhen, 1000 Genomes Project, dbSNP, and SwissProt annotations among others), and selection of high confidence variants; (*5.*) identification of potentially pathogenic variants and VUCS based on their confidence, effect prediction, and population frequencies.

was prepared according to the PacBio's 5 or 10 Kb Template Preparation procedures. Pooled amplicons were sequenced on 5 SMRT cells on the PacBio RSII system with the P6 sequencing

chemistry. Data collected from 360-minute movie-time was preprocessed using the standard primary analysis tools (Figure 1).

2.4.2.3. Sequence Alignment and Variant Calling

We used the RS_Resequencing protocol from SMRT Analysis Suite v.2.3 to align long-reads against the human reference genome, downloaded from the Genome Reference Consortium version 37 patch 13 (GRCh37.p13). Samples were de-multiplexed into individual files (H5 and BAM formats) using known barcode sequences and a minimum barcode identity score of 22 (Figure 1). Alignments were filtered to contain mapped reads with a mapping quality threshold of 30 Phred score using Samtools v.1.2 (Supp. Table S2). Read coverage and targeted PCR statistics for LR-PCR amplicons were computed with BedTools v.2.25.0, and PicardTools v.1.1.40 (Supp. Table S2; Supp. Table S3; Supp. Table S4). Variant calling was performed using Quiver (allowing for diploid calling and maximum coverage of 10,000). Variants with a Quiver confidence score lower than 40 were filtered out from downstream analysis (Figure 1).

2.4.2.4. Reconstruction of Allelic Sequences

PKD1 and *PKD2* allelic sequences were reconstructed by using the Long Amplicon Analysis (LAA) available in SMRT Analysis Suite 2.3. Only reads longer than 3,000 base pairs (bp) and average signal to noise ratio of 3 were used for the reconstruction (Figure 1). Based on this reference-free subread (full-length and non-full-length reads) clustering, chimeric sequences were identified and comprehend ≤0.85% (6,288/738,822) of subreads that were subsequently removed from the analysis. Allelic sequences of *PKD1* and *PKD2* were aligned to the human reference genome GRCh37.p13 using BLASR (Chaisson and Tesler, 2012), and reporting a single best-scoring alignment. Variants were extracted by comparison between the reconstructed alleles and the human reference sequence with the Variant Description Extractor from the Mutalyzer Suite 2.0.21 (Vis et al., 2015).

2.4.2.5. Loss of Heterozygosity (LoH) Analysis

Loss of heterozygosity (LoH) for each amplified fragment was assessed to identify patients with potential large deletions for *PKD1*. We first identified heterozygous substitutions within the amplified fragments with a variant frequency between 25% and 75%. Amplified fragments with zero heterozygous substitutions were identified as LoH. Large deletions produce multiple LR-PCR fragments dropouts, and were identified by the detection of consecutive LoH fragments.

The detection of consecutive LoH fragments was not a direct identification of large deletions *per se*, rather than an indication of the presence of large deletions in the amplified LR-PCR fragment regions. Identified LoH regions were then compared with large deletions detected by MLPA.

2.4.3. ADPKD Variant Nomenclature and Genotyping

Variant descriptions were standardized to concord with HGVS guidelines (den Dunnen et al., 2016), using the Mutalyzer Name Checker tool (Wildeman et al., 2008). Genomic HGVS descriptions were converted to coding notations using the Position Converter from Mutalyzer (Wildeman et al., 2008). Only changes in RefSeq annotated canonical transcripts for PKD1 (NM_001009944.2) and PKD2 (NM_000297.3) were further analyzed. HGVS descriptions of deletion-insertions (delins) were manually inspected to avoid variant redundancies and undesired clustering of neighboring independent events (Figure 1; Supp. Table S5). Then, standardized variants were annotated using the Variant Effect Predictor (VEP), from Ensembl tools v.83 (McLaren et al., 2010), with additional parameters "-everything", and "-refseq" (Figure 1). All variant annotations reported by VEP are fully disclosed in the raw VCF files (EGAS00001002106). Variant frequency and coverage were used to filter low confidence variants by applying thresholds for: (a.) sequencing depth of ≥50x subreads and ≥15x reads which ensures a sufficient control over the SMRT sequencing random error rate (1% mismatches and 13% indels); (b.) minimum variant frequency of 10% for substitutions and 15% for deletions and insertions (Figure 1). For interpreting insertion and deletion frequencies, neighboring bases were also examined. The selection of strong pathogenic variant candidates or variants of unknown clinical significance (VUCS) was based on the following criteria: (1.) high predicted effect on the coding sequence or splice site region (e.g. missense, in-frame indels, frameshifts, and splice site acceptor or donor variants); (2.) population frequency in the 1000 genomes project <1%; (3.) unique occurrence (1/19) (~5%) in the patient cohort since, in ADPKD, no single disease causing variant accounts for more than 2% of affected families (Harris and Rossetti, 2010), or more than 1.7% of ADPKD reported cases in the ADPKD database (PKDB) (http://pkdb.mayo.edu/; accessed version 3.1) (Gout et al., 2007) (Figure 1).

2.4.4. Clinical diagnostics pipeline for ADPKD genotyping

2.4.4.1. Sanger Sequencing

The current diagnostics pipeline for ADPKD genotyping, including Sanger sequencing and MLPA, used a different set of LR-PCR primers to target the duplicated part of *PKD1* (exons 1 to 32) (Supp. Table S6). The non-duplicated region of *PKD1* (exons 33 to 46), and *PKD2* regions (exons 1 to 5) were amplified using targeted standard PCR reactions (Supp. Figure S1), with 100 ng of input genomic DNA with M13 tail primers. The nested and standard PCR amplicons were designed to cover the complete coding regions and splice sites with at least 20 bp of flanking intronic sequences (Supp. Table S7; Supp. Table S8). The duplicated part of *PKD1*, which includes exons 1 to 32, was amplified using 4 different LR-PCR fragments that covered exons 1, 2-13, 14-21, and 22-32, respectively (Supp. Table S6). LR-PCR amplification was performed using Thermo Scientific 2x Extensor Long Range PCR Master Mix on 50 ng of DNA. Then, a nested-PCR was carried out on 4 µl 100-250x of diluted product to obtain the final Sanger sequencing fragments. The nested-PCR primers with an M13 tail were used to amplify the coding region including 5 to 20 bp of intronic sequences (Supp. Figure S1). Large exons such as exon 5, 10, 11, 15, and 23 were amplified using overlapping nested PCR products, although 10 bases of exon 15 (c.6503-6514) were not covered. Nested PCR and standard PCR of the non-duplicated part of *PKD1*, and *PKD2*, was carried out in a final volume of 15 µl in GoTaq Colorless Taq Reaction buffer with 0.6 U of Taq DNA-polymerase (Promega) at a final concentration 5 pM for each primer, 200 µM of each dNTP. After a hot start at 95 °C, a denaturation was performed for 5 min at 95 °C, followed by 35 cycles of 45 sec at 94°C, 45 sec at 60°C, and 30 sec at 72 °C. The final extension was of 5 min at 72 °C in a T-Professional Thermocycler (Biometra, Westburg). All liquid handling steps were automated using the SciClone (ALH-HV96 pipetting station, Perkin Elmer) or Biomek FX workstation (Beckman). PCR products (20 to 50 ng) were purified using an Ampure XP PCR purification kit and sequenced using BigDye Terminator v3.1 sequencing reactions (Applied Biosystems) with PAGE purified –21M13 or M13REV sequencing primer. The excess of dye terminatiors was removed by gel filtration using the Edge Biosystem Dye Terminator Removal (DTR) with a 96 wells plate. After electrophoresis on an ABI Prism 3730 (XL) DNA analyzer (Life technologies, Applied Biosystems) data processing was automated using SeqPatient software (Sequence Pilot, JSI medical systems, GMbH).

2.4.4.2. Multiplex Ligation-dependent Probe Amplification (MLPA)

To detect large deletions and duplications, two commercially available MLPA kits (P351-B2 and P352-C1; MRC-Holland, Amsterdam, The Netherlands) were used following manufacturer's protocols and manuals.

2.4.5. Comparative Analysis of SMRT sequencing and current ADPKD diagnostic assay

The overlap between identified variants based on PacBio and Sanger sequencing data was achieved by assessing identical standardized HGVS descriptions. Only variants with predicted effects on coding DNA or splice site regions were considered (Supp. Table S5). PacBio and Sanger variants were manually inspected to detect overlapping variants with discordant descriptions between the two datasets. To facilitate interpretation, each unique variant was further annotated with its PKDB clinical significance, Single Nucleotide Polymorphism database version 144 (dbSNP) identifier, and the number of patients where it was detected in the cohort. Surrounding bases were evaluated to identify and remove potential sequencing artifacts occurring in homopolymer stretches. Finally, variants were considered as high-confident variants if previously reported in PKDB or dbSNP, showed strong PacBio sequencing evidence of being present, or detected in any patient by both Sanger and PacBio.

2.4.6. Short-read loss of power for known *PKD1* pathogenic variants in WGS and WES

Previously known pathogenic variants for *PKD1* gene were obtained from PKDB. Only variants that were classified as "definitely pathogenic" were selected for further analysis. Large deletions (few hundred bp to several Kbp long) were excluded from the analysis as they are not usually detected with common variant calling algorithms. For the genomic position of each pathogenic variant, sequencing depth was extracted from 9 publicly available WGS and whole-exome sequencing (WES) datasets (Sun et al., 2015). In addition, we included the sequencing depth of 9 randomly selected libraries from the study of Rossetti *et al*, 2012, in which the authors used a similar strategy based on LR-PCR followed by short-read sequencing. Each library represents an equimolar pool of DNA from 4 different patient samples that were not possible to further de-multiplex because individuals were not barcoded. Variant positions with low sequencing depth (< 8 reads, or < 32 for the short-read LR-PCR approach) were marked as inaccessible positions of clinical significance using BedTools v2.25.0. Finally, variant positions were classified into three categories based on the number of individuals with poor coverage at each position: (*1.*)

variants with sufficient coverage in all 9 individuals; (*2.*) variants reported inaccessible in 2–4 individuals; and (*3.*) variants reported inaccessible in 5 or more individuals.

2.4.7. Data Availability

Sequencing data and alignments in BAM format can be accessed through the European Genome-phenome Archive (EGA), as well as raw variants in VCF file format, under the EGA study identifier EGAS00001002106. Coding or splice site variants were also uploaded to the Leiden Open Variation Database (LOVD). Description and examples of custom scripts used in this manuscript are accessible upon request from a local GitLab repository.

2.5. Results

2.5.1. Targeted sequencing of ADPKD genes

Direct sequencing of LR-PCR fragments (designed to specifically and uniquely amplify *PKD1*, and *PKD2* gene regions) (Supp. Figure S1; Supp. Table S1) was performed to evaluate the utility of long-read sequencing in resolving ADPKD for molecular diagnostics. All *PKD1* and *PKD2* exons (including the duplicated part of *PKD1*, as well as 20bps of flanking intron regions) from 19 ADPKD patients could be completely covered using long-reads, sequenced on the PacBio RSII platform (Figure 2; Supp. Figure S2). Most of long-reads (94.4%) were uniquely mapped to *PKD1* and *PKD2* (Supp. Table S2). Reads originating from residual off-target amplification (5.6%; Supp. Table S2) introduced during the LR-PCR steps were identified, and discriminated, by their unique alignment to the *PKD1* pseudogenes (Figure 2; Supp. Table S2). All *PKD1* and *PKD2* protein-coding and flanking intron sequences (±20 bp) were covered at average sequencing depth ≥421x (min. ≥19x; max. 1528x), with ≥97.36% of bases over ≥30x, which was well above the applied threshold of ≥15x reads (Supp. Table S2; Supp. Table S4). Amplicons that cover the first and last exons of *PKD1* were underrepresented when compared to other LR-PCR fragments, with a total of ≥593 average reads (min. ≥300; max 1580) and ≥87 (min. ≥35; max 153) for *PKD1* fragments A and E, respectively (Figure 2B; Supp. Table S3). The usually difficult to sequence first exons of both, *PKD1* and *PKD2* genes were covered, on average ≥55x (min. ≥24x; max. 91x) and ≥71x (min. ≥43x; max. 111x), respectively (Supp. Table S4). Most of sequenced reads (>99.9%) were uniquely mapped to *PKD1* and *PKD2* (Figure 2B; Supp. Figure S2).

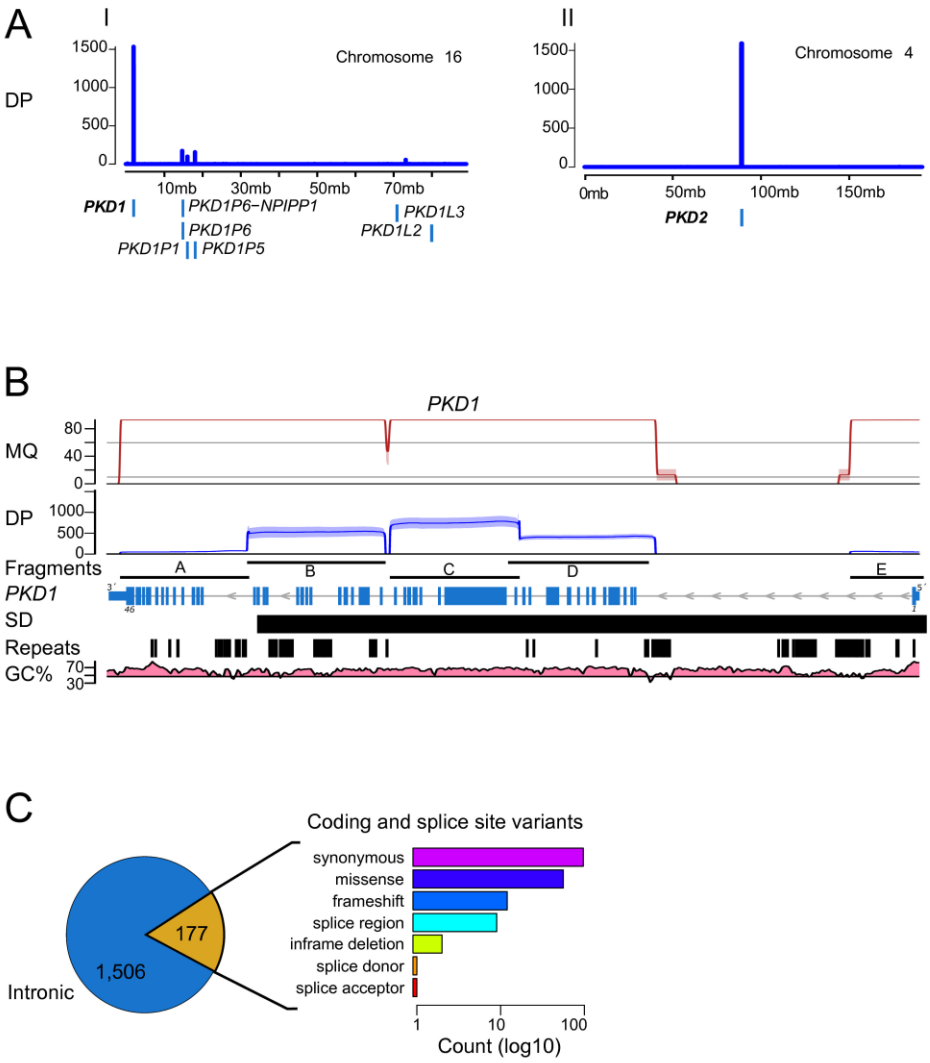


Figure 2: SMRT sequencing and variant calling of LR-PCR amplicons. (*A*) Sequencing depth (DP; in number of reads) of the alignments to chromosome 16 and chromosome 4. Number of uniquely aligned reads (y axis, blue line) sequenced with PacBio that mapped to *PKD1* and *PKD2*. Off-target amplification is discriminated from the main *PKD1* gene sequences showing alignments to pseudogene homologous sequences at proximal loci (e.g. *PKD1P1*, *PKD1P5*, *PKD1P6*) (blue boxes). (*B*) Mapping quality (MQ; in Phred quality scores; values >90 were scaled down for visualization purposes), and sequencing depth (DP; in number of reads) of uniquely aligned molecules to *PKD1* (NM_001009944.2) for the 5 LR-PCR fragments amplified. Mapping quality of alignments with even coverage distribution along the amplified fragments (Fragments), including regions with segmental duplications (SD), repetitive elements (Repeats) and high GC-content (GC%). Despite fragments A and E show lower coverage, compared to the average sequencing depth of ≥421x (min. ≥19x; max. 1528x), they had sufficient coverage for variant calling within the exon regions, including the first exons of *PKD1*, with average coverage of ≥55x (min. ≥24x; max. 91x) (Supp. Table S4). *(C)* We detected 1,506 intron variants (blue) and 177 coding or splice site variants (yellow). The predicted transcript effects of coding and splice site variants were quantified (bar chart) as log10 count (x axis).

2.5.2. Sensitive detection of ADPKD small variants

PKD1 is known to be a highly polymorphic gene with many variants reported in addition to the disease causing or pathogenic variants (Gout et al., 2007). Hence, the required sensitivity to resolve *PKD1* was achieved by the combination of variant calling using Quiver and the reconstruction of amplified allelic sequences. Overall, we identified 1,683 variants (404 SNPs) across 19 ADPKD patients, from which 177 variants (119 SNPs) were located in coding or splice site regions (Figure 2C). Variants were distributed along *PKD1* (Supp. Figure S3A) including regions with large segmental duplications and high GC-content. The mismatch rate of PacBio data was empirically assessed based on average frequency of mismatches at each position. We observed an average of 1.2% mismatch rate across the entire *PKD1* gene (Supp. Figure S3A). This correlates with the random sequencing mismatch rate of 1% for PacBio and thus, the applied minimum frequency threshold of 10% for substitutions is well above the observed noise introduced by random PacBio errors.

2.5.3. Large deletions in *PKD1*

Detection of allele dropouts and large deletions in *PKD1* was assessed by performing a loss of heterozygosity (LoH) analysis for each of the amplified regions (Figure 3). We identified 17 LR-PCR fragments with LoH among all 19 patients sequenced. Most of these (10) corresponding to LoH regions identified in fragment E (Figure 3). Only 2 patients showed consecutive LoH fragments indicating the presence of large deletions spanning between two or more LR-PCR fragments. These consecutive LoH fragments are not a direct identification of the deletions *per se* but an indication of the presence of large deletions in the amplified region. The two patients that showed two or more consecutive fragments with LoH (Figure 3) were in concordance with large deletions identified by MLPA as pathogenic variants in the same ADPKD patients. A deletion of ≥1,543 bps (c.(2097+1_2098-1)_3640del; exons 11-15) and a deletion of ≥9,108 bps (c.(287+1_288-1)_(9397+1_9398-1)del; exons 3-26) were detected by MLPA for patient sample 7 and 13, respectively (Figure 3). With the current experimental design, however, the exact location of the breakpoints for each large deletion could not be determined with either method.

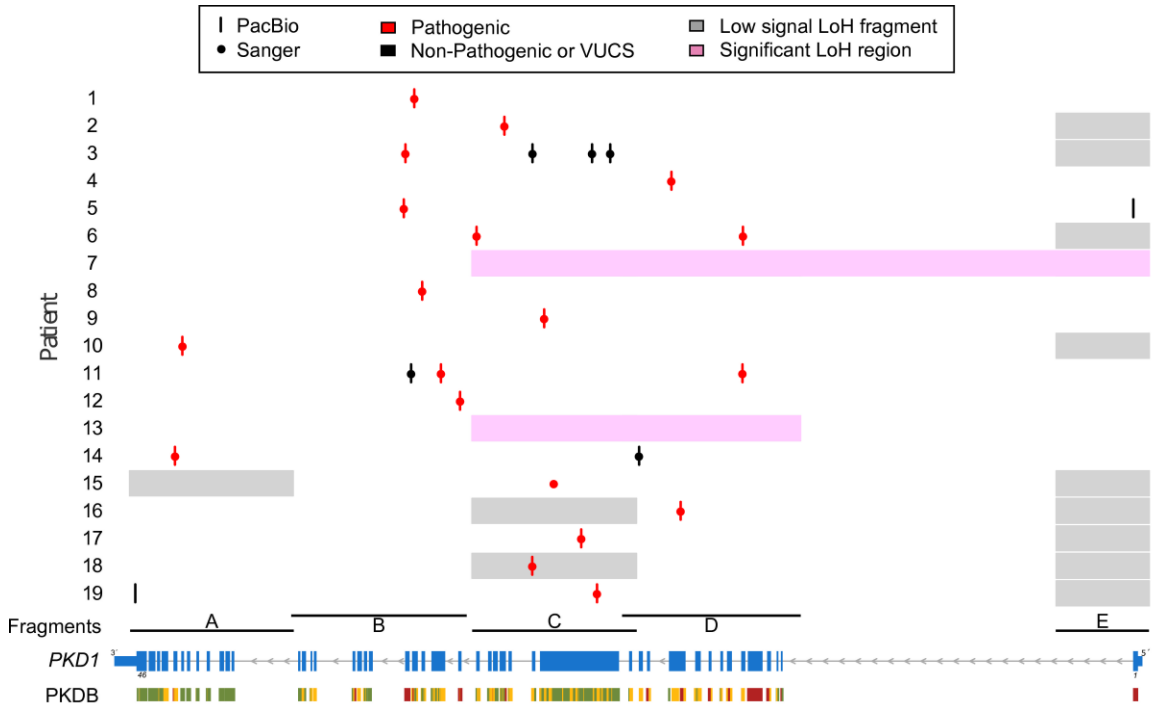


Figure 3: Comparison of long-read detected pathogenic variants or VUCS, uniquely identified per patient (y axis), with the screening results for the *PKD1* gene locus (x axis; NM_001009944.2). Most of pathogenic variants (red) could be confirmed by our long-read strategy (red bars) with high sensitivity for *PKD1*. Only a single insertion could not be confirmed for patient 16. Other identified non-pathogenic variants or VUCS are shown as black bars and dots for PacBio and Sanger, respectively. The LoH analysis performed (pink or grey boxes) support the presence of the 2 large deletions also reported by MLPA (pink boxes). LoH regions are not a direct identification of large deletions but a clear indication of their presence within the amplified LR-PCR fragments.

2.5.4. Comparative Analysis between SMRT-Seq and the ADPKD diagnostic assay

The evaluation of 167 coding or splice site variants identified by standard ADPKD diagnostic assay, showed that 159 out of 167 were correctly detected by PacBio (Supp. Figure S3C). The overall observed sensitivity and specificity in detecting coding variants was of 95.2% (159/167) and 88.8% (159/179), respectively. Eight variants were solely detected by Sanger (Supp. Figure S3A: crosses), from which, despite of high sequencing depth, the majority (6/8) had low number of reads supporting the presence of these variants in PacBio sequencing data with variant frequency below the applied frequency thresholds (Supp. Figure S3B: yellow dots). The remaining 2/8 variants (Supp. Figure S3B: red dots) constitute a pathogenic insertion (c.6223_6224insTT) and one polymorphic substitution (c.12630T>C) (Supp. Table S9; Supp. Figure S4).

From 179 variants detected by PacBio, 20 were solely identified by PacBio (Supp. Figure S3C; Supp. Table S9). Of these, 17/20 were high-confident variants not detected in Sanger. The remaining 3/20 were low confidence variants from the reconstruction of allelic sequences for the variant c.6657_6671del (Supp. Table S9).

Table 1: Uniquely identified pathogenic variants or variants of unknown clinical significance, identified by PacBio sequencing.

Patient	Genomic position		Exon	c. notation	p. notation	SNP ID	Freq (%)	Depth	PolyPhen	VEP impact	Comparison with Sanger
1	**chr16**	**2,152,543**	**25**	**c.9034_9039del**	**p.(Thr3012_Ser3013del)**		**40.9**	**314**		**MODERATE**	**Overlap**
2	**chr16**	**2,156,674**	**18**	**c.7214G>T**	**p.(Trp2405Leu)**		**29.4**	**666**	**probably damaging (0.983)**	**MODERATE**	**Overlap**
3	chr16	2,161,525	15	c.3643C>G	p.(Leu1215Val)	rs144338515	49.2	576	possibly damaging (0.899)	MODERATE	Overlap
3	chr16	2,160,693	15	c.4475G>C	p.(Arg1492Pro)		32.8	563	possibly damaging (0.665)	MODERATE	Overlap
3	chr16	2,157,963	16	c.6986G>A	p.(Arg2329Gln)	rs575211353	43.3	538	benign (0.37)	MODERATE	Overlap
3	**chr16**	**2,152,134**	**26**	**c.9324del**	**p.(Ile3109SerfsTer207)**	**rs780284643**	**26**	**302**		**HIGH**	**Overlap**
4	**chr16**	**2,164,333**	**11**	**c.2681_2690del**	**p.(Phe894Ter)**		**53.7**	**143**		**HIGH**	**Overlap**
5	chr16	2,185,509	1	c.182C>T	p.(Pro61Leu)		28.8	43	benign (0.119)	MODERATE	PacBio
5	**chr16**	**2,152,061**	**26**	**c.9397+1G>A**			**23.3**	**330**		**HIGH**	**Overlap**
6	**chr16**	**2,167,614**	**6**	**c.1261C>T**	**p.(Arg421Cys)**		**38.7**	**273**	**possibly damaging (0.836)**	**MODERATE**	**Overlap**
6	**chr16**	**2,155,399**	**21**	**c.7940C>T**	**p.(Thr2647Met)**	**rs748496650**	**44**	**357**	**probably damaging (1)**	**MODERATE**	**Overlap**
7	**chr16**	**2,161,527**		**c.(2097+1_2098-1)_3640del**							**PacBio**
8	**chr16**	**2,152,903**	**24**	**c.8859dup**	**p.(Glu2954Ter)**		**44.1**	**583**		**HIGH**	**Overlap**
9	**chr16**	**2,158,496**	**15**	**c.6657_6671del**	**p.(Arg2220_Pro2224del)**		**41.5**	**466**		**MODERATE**	**Overlap**
10	**chr16**	**2,141,910**	**40**	**c.11412-3C>A**			**27.7**	20		**LOW**	**Overlap**
11	**chr16**	**2,167,589**	**6**	**c.1286G>T**	**p.(Trp429Leu)**		**32.7**	**313**	**probably damaging (0.999)**	**MODERATE**	**Overlap**
11	**chr16**	**2,153,765**	**23**	**c.8293C>T**	**p.(Arg2765Cys)**	**CM092156 rs144979397**	**41.1**	**572**	**probably damaging (0.988)**	**MODERATE**	**Overlap**
11	chr16	2,152,396	25	c.9187C>T	p.(Arg3063Cys)	rs145906459	36.1	557	benign (0.39)	MODERATE	Overlap
12	**chr16**	**2,154,643**	**21**	**c.8017-2_8017-1del**			**48.5**	**527**		**HIGH**	**Overlap**
13	**chr16**	**2,152,062**		**c.(287+1_288-1)_(9397+1_9398-1)del**							**PacBio**
14	**chr16**	**2,141,581**	**42**	**c.11554del**	**p.(Leu3852TrpfsTer93)**	**rs724159823**	**41.3**	**46**		**HIGH**	**Overlap**
14	chr16	2,162,850	13	c.3100A>G	p.(Asn1034Asp)	rs369180760	36.5	321	benign (0.098)	MODERATE	Overlap
15	**chr16**	**2,158,944**	**15**	**c.6223_6224insTT**	**p.(Arg2075LeufsTer42)**		**43.3**	**1,203**		**HIGH**	**Sanger**
16	**chr16**	**2,164,754**	**11**	**c.2269del**	**p.(Gln757SerfsTer28)**	**rs775710328**	**28.4**	**519**		**HIGH**	**Overlap**
17	**chr16**	**2,160,198**	**15**	**c.4968_4969delinsC**	**p.(Arg1657GlyfsTer65)**		**41.2**	**1,120**		**HIGH**	**Overlap**
18	**chr16**	**2,157,954**	**16**	**c.6994_7000dup**	**p.(Ala2332GlyfsTer90)**		**23.4**	**913**		**HIGH**	**Overlap**
19	chr16	2,139,750	46	c.12890A>G	p.(Lys4297Arg)	rs758833703	14.1	46	benign (0.07)	MODERATE	PacBio
19	**chr16**	**2,160,919**	**15**	**c.4248dup**	**p.(Gly1417TrpfsTer14)**		**24.8**	**979**		**HIGH**	**Overlap**

Sanger detected pathogenic variants are shown in **bold**. PacBio variants were filtered by coding sequence predicted effects (frameshifts, missense, in-frame deletions, and splicing variants), as well as DP>15 and >50 subreads, and variant frequency (>10% for substitutions, and >15% for insertions and deletions) (RefSeq NM_001009944.2). Additional information of each variant including SIFT classification, and 1000G frequencies among other annotations can be obtained from the VCF files uploaded to EGA with accession number EGAS00001002106.

The sensitivity assessment for *PKD1* pathogenic variants was performed by comparing the list of potentially pathogenic variants and VUCS, uniquely identified by our direct long-read sequencing approach, with the results from the standard ADPKD diagnostic assay. Although we expected a single dominant pathogenic variant per patient, two of the patients had a combination of two pathogenic variants resulting in 21 *PKD1* pathogenic variants. We identified 20 out of 21 pathogenic variants (95.2%) in addition to 7 VUCS from which 2 were uniquely detected by PacBio (Table 1; Figure 3). Only a single pathogenic insertion (c.6223_6224insTT) was missed by PacBio variant calling despite of sufficient read support (43.3% variant frequency; read depth

1,203) (Table 1). In summary, 18 out of 19 ADPKD patients could be resolved by our method (Figure 3). This provided a diagnosis for 94.7% of the patients, resulting in the correct detection of all *PKD1* substitutions, single-nucleotide deletions, large deletions, one deletion-insertion, and 3 out of 4 insertions or duplications (Table 1).

2.5.5. Loss of *PKD1* diagnostic power in short-read (Illumina) NGS

The potential loss of diagnostic power when resolving *PKD1* by short-read NGS was evaluated based on 797 pathogenic variants that were previously reported and validated, and are publicly available in PKDB. The repetitive nature of *PKD1* gene hampers proper alignment of short Illumina NGS reads (Supp. Figure S5). Over 12% of the reported pathogenic variants would have been missed in WGS and WES data purely due to poor sequencing depth (Supp. Figure S6). In comparison, other short-read approaches based on LR-PCR enrichment show lower percentage (1.3%) of reported pathogenic variants that would have been missed because of low sequencing depth. However, this approach required very high sequencing depth which can be appreciated from the observed high variability in coverage ranging from <8x to >30,000x (Supp. Figure S6). Moreover, several exonic regions may be expected to be missed in many samples irrespective of the short-read sequencing strategy used (Supp. Figure S6).

2.6. Discussion

Accurate diagnosis is a difficult task when performed in complex genetic regions such as *PKD1* (Rossetti et al., 2007; Tan et al., 2009; Qi et al., 2013). To facilitate the diagnosis, we have developed and applied a new methodology using direct long-read sequencing of amplified LR-PCR fragments on PacBio. Because of the repetitive nature of *PKD1*, current diagnostics is performed by Sanger sequencing using LR-PCR fragments generated for approximately two thirds of the *PKD1* gene that serve as a template for the exon specific nested-PCR amplification. In contrast, in this study we directly sequenced all LR-PCR fragments amplified from the duplicated and unique parts of *PKD1* gene as well as *PKD2*. On top of reducing the PCR amplification steps required and limiting the implicit PCR artifacts, single molecule sequencing improves sequence alignments and aids in discriminating between homologous or repeated sequences, such as *PKD1* pseudogenes. This provides a cleaner dataset for variant calling, free of chimeric (0.85%) and pseudogene (5.6%; Supp. Table S2) reads that are introduced by the LR-PCR amplification. Finally, by using this approach, we identified 20 out of 21 (95.2%) *PKD1*

disease causing variants diagnosed by Sanger sequencing or MLPA, providing a correct diagnosis for 18 out of 19 ADPKD patients (94.7%) with at least one pathogenic variant in *PKD1*.

In comparison to current ADPKD diagnostic assays, based on Sanger sequencing and MLPA, we show that direct long-read sequencing can aid in resolving *PKD1* for ADPKD diagnostics. Longer sequencing reads discriminate between *PKD1* and pseudogenes (Figure 2A), and improve the mapping quality of *PKD1* (Figure 2B). The improved mappability, reduced the interference of homologous sequences, high GC-content, or repetitive elements for ADPKD diagnosis (Qi et al., 2013). This allowed us to develop a long-read based sequencing assay for detecting a broad spectrum of variants, from SNVs to large deletions (Table 1). In contrast, Sanger sequencing is very labor-intensive and requires many phases of overlapping PCR amplification steps prior to sequencing, including LR-PCR and nested-PCR. Despite the amplification of *PKD1* being based on unique PCR primers, these are of limited number for *PKD1* and have been shown to produce residual amplification of homologous regions that would still interfere with the aggregated signal of Sanger sequencing (Rossetti et al., 2012; Tan et al., 2014). Based on our approach, we confirmed the presence of residual amplification of *PKD1* pseudogenes, introduced by the LR-PCR (5.6%) (Figure 2A; Supp. Table S2). This, most likely led to the identification of 24 false-positive or false-negative variants detected by Sanger (Supp. Table S9; Supp. Figure S4). One of the major drawbacks of our method, however, is the noise associated with PacBio sequencing, and the sophisticated algorithms required to overcome it. This noise is likely to be the cause of most of 324 homopolymer deletion artifacts that were solely identified by PacBio (Supp. Table S10). In addition, this noise was the most likely cause of the single pathogenic insertion that was missed despite ample sequencing depth. However, based on a recent release of the new circular-consensus calling algorithm for PacBio sequencing data (www.pacb.com: *"An improved circular consensus algorithm with an application to detect HIV-1 drug-resistance associated with mutations (DRAMS)"*), we expect that calling of true homopolymer-associated variants will be significantly improved.

In recent years, several attempts have been made to replace the standard ADPKD diagnostics by NGS approaches that would improve the screening of *PKD1* gene (Rossetti et al., 2012; Qi et al., 2013; Tan et al., 2014; Trujillano et al., 2014; Eisenberger et al., 2015; Mallawaarachchi et al.,

2016). These screenings were based on analyzing WGS or WES data (Qi et al., 2013; Mallawaarachchi et al., 2016), on the enrichment of *PKD1* using LR-PCR (Rossetti et al., 2012; Tan et al., 2014), or the hybridization-capture of *PKD1* (Trujillano et al., 2014; Eisenberger et al., 2015). Two of these studies were performed on short-read NGS using targeted enrichment of *PKD1* or *PKD2* genes by LR-PCR (Rossetti et al., 2012; Tan et al., 2014). In both studies, the use of short reads was the source of difficulties associated with misalignments and lack of sufficient coverage, such as the *PKD1* exon 1 region (Tan et al., 2014), as well as false positive (10%) and false negative variant calls (5%) (Rossetti et al., 2012). We show that these challenges were mitigated with long-read sequencing which provided 100% coverage >10x (min. >19x; avg. >421x; max. 1528x) for all *PKD1* and *PKD2* exons and flanking intron regions (±20 bp) (Supp. Table S2; Supp. Table S4), including 100% of *PKD1* exon 1 at average coverage of >55x (±20 bp of flanking intron regions included) (Supp. Table S4). Other WES-based strategies were reported to resolve only 50% of true positive variants in the duplicated regions of *PKD1* (Qi et al., 2013). It was argued that increasing the sequencing depth was insufficient to overcome the limitations and pitfalls of short-read NGS approaches (Qi et al., 2013; Eisenberger et al., 2015). Similar to these short-read NGS strategies (Rossetti et al., 2012; Qi et al., 2013; Tan et al., 2014; Trujillano et al., 2014; Eisenberger et al., 2015; Mallawaarachchi et al., 2016), our targeted approach combined with multiplexed sequencing can further accelerate ADPKD diagnostics, compared to the labor-intensive Sanger sequencing (Rossetti et al., 2012; Tan et al., 2014). Despite the rather limited sample size, sufficient numbers were included in this study for a methodology evaluation. However, future studies including larger cohorts would be needed to reliably implement the proposed methodology into the clinic. In addition, our method can benefit from recent advancements in library preparation methods with minimal or no amplification, such as single-strand adaptor-ligation (Karlsson et al., 2015), which would eliminate most of biases introduced during LR-PCR amplification steps (Schirmer et al., 2015; Hestand et al., 2016; Laver et al., 2016). Overall, our method provides high sensitivity in identifying *PKD1* genetic variants when compared to the standard ADPKD diagnostic assay and showed an added value to become a reliable alternative. In addition, the method presented here is comparable to other Illumina short-read NGS-based approaches. However, further studies including a larger cohort may be required to decipher the true diagnostic power of our approach compared to that of

standard ADPKD diagnostic assays using Sanger and MLPA, and to Illumina short-read NGS-based methods.

In conclusion, we showed that direct sequencing of LR-PCR fragments for the screening of ADPKD patients in a single diagnostic-test application is now possible. Accurate screening of *PKD1* with high sensitivity and low interference of homologous sequences constitutes a clear example. This method is highly valuable for a diagnostic setting, as it increases the resolution power of clinically relevant but difficult to sequence or to resolve genomic regions.

2.7. Acknowledgements

The authors declare no conflict of interest.

This study was performed within the scope of the iMODE-CKD Initial Training Network (ITN) (Clinical and system-omics for the identification of the Molecular DEterminants of established Chronic Kidney Disease). The research leading to these results has received funding from the European Union's Seventh Framework Programme FP7/2007-2013 under grant agreement FP7-PEOPLE-2013-ITN-608332. (www.imodeckd.org)

Data obtained from Rossetti *et al*, 2012 was kindly provided by Peter C. Harris and Christina M. Heyer (Mayo Clinic College of Medicine).

Patient samples were kindly provided by the DIPAK Consortium, an inter-university collaboration in the Netherlands, established to study autosomal dominant polycystic kidney disease and to develop rational treatment strategies for this disease (www.nierstichting.nl/dipak). Principal investigators are (in alphabetical order): J. P. H. Drenth (Department of Gastroenterology and Hepatology, Radboud University Medical Center Nijmegen), J. W. de Fijter (Department of Nephrology, Leiden University Medical Center), R. T. Gansevoort (Department of Nephrology, University Medical Center Groningen), D. J. M. Peters (Department of Human Genetics, Leiden University Medical Center), J. Wetzels (Department of Nephrology, Radboud University Medical Center Nijmegen), and R. Zietse (Department of Internal Medicine, Erasmus Medical Center Rotterdam).

We acknowledge Joost P. Schanstra for his continued support (Inserm; Université Toulouse III Paul Sabatier, Institut de Médicine Moléculaire de Rangueil).

2.8. References

Barua M, Cil O, Paterson AD, Wang K, He N, Dicks E, Parfrey P, Pei Y. 2009. Family History of Renal Disease Severity Predicts the Mutated Gene in ADPKD. J Am Soc Nephrol 20:1833–1838.

Chaisson MJ, Tesler G. 2012. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. BMC Bioinformatics 13:238.

Chaisson MJP, Huddleston J, Dennis MY, Sudmant PH, Malig M, Hormozdiari F, Antonacci F, Surti U, Sandstrom R, Boitano M, Landolin JM, Stamatoyannopoulos JA, et al. 2015. Resolving the complexity of the human genome using single-molecule sequencing. Nature 517:608–611.

Chang F, Li MM. 2013. Clinical application of amplicon-based next-generation sequencing in cancer. Cancer Genet 206:413–419.

Codina-Solà M, Rodríguez-Santiago B, Homs A, Santoyo J, Rigau M, Aznar-Laín G, Campo M del, Gener B, Gabau E, Botella MP, Gutiérrez-Arumí A, Antiñolo G, et al. 2015. Integrated analysis of whole-exome sequencing and transcriptome profiling in males with autism spectrum disorders. Mol Autism 6:21.

Dewey FE, Grove ME, Pan C, Goldstein BA, Bernstein JA, Chaib H, Merker JD, Goldfeder RL, Enns GM, David SP, Pakdaman N, Ormond KE, et al. 2014. Clinical Interpretation and Implications of Whole-Genome Sequencing. JAMA 311:1035.

Dunnen JT den, Dalgleish R, Maglott DR, Hart RK, Greenblatt MS, McGowan-Jordan J, Roux A-F, Smith T, Antonarakis SE, Taschner PEM. 2016. HGVS Recommendations for the Description of Sequence Variants: 2016 Update. Hum Mutat 37:564–569.

Eisenberger T, Decker C, Hiersche M, Hamann RC, Decker E, Neuber S, Frank V, Bolz HJ, Fehrenbach H, Pape L, Toenshoff B, Mache C, et al. 2015. An Efficient and Comprehensive Strategy for Genetic Diagnostics of Polycystic Kidney Disease. PLOS ONE 10:e0116680.

Gout AM, Martin NC, Brown AF, Ravine D. 2007. PKDB: Polycystic Kidney Disease Mutation Database—a gene variant database for autosomal dominant polycystic kidney disease. Hum Mutat 28:654–659.

Guo X, Zheng S, Dang H, Pace RG, Stonebraker JR, Jones CD, Boellmann F, Yuan G, Haridass P, Fedrigo O, Corcoran DL, Seibold MA, et al. 2013. Genome Reference and Sequence Variation in the Large Repetitive Central Exon of Human MUC5AC. Am J Respir Cell Mol Biol 50:223–232.

Harris PC, Rossetti S. 2010. Molecular diagnostics for autosomal dominant polycystic kidney disease. Nat Rev Nephrol 6:197–206.

Hestand MS, Houdt JV, Cristofoli F, Vermeesch JR. 2016. Polymerase specific error rates and profiles identified by single molecule sequencing. Mutat Res Mol Mech Mutagen 784–785:39–45.

Karlsson K, Sahlin E, Iwarsson E, Westgren M, Nordenskjöld M, Linnarsson S. 2015. Amplification-free sequencing of cell-free DNA for prenatal non-invasive diagnosis of chromosomal aberrations. Genomics 105:150–158.

LaDuca H, Stuenkel AJ, Dolinsky JS, Keiles S, Tandy S, Pesaran T, Chen E, Gau C-L, Palmaer E, Shoaepour K, Shah D, Speare V, et al. 2014. Utilization of multigene panels in hereditary cancer predisposition testing: analysis of more than 2,000 patients. Genet Med 16:830–837.

Laver TW, Caswell RC, Moore KA, Poschmann J, Johnson MB, Owens MM, Ellard S, Paszkiewicz KH, Weedon MN. 2016. Pitfalls of haplotype phasing from amplicon-based long-read sequencing. Sci Rep 6:21746.

Lee H, Schatz MC. 2012. Genomic dark matter: the reliability of short read mapping illustrated by the genome mappability score. Bioinformatics 28:2097–2105.

Ligt J de, Willemsen MH, Bon BWM van, Kleefstra T, Yntema HG, Kroes T, Vulto-van Silfhout AT, Koolen DA, Vries P de, Gilissen C, Rosario M del, Hoischen A, et al. 2012. Diagnostic Exome Sequencing in Persons with Severe Intellectual Disability. N Engl J Med 367:1921–1929.

Loomis EW, Eid JS, Peluso P, Yin J, Hickey L, Rank D, McCalmon S, Hagerman RJ, Tassone F, Hagerman PJ. 2013. Sequencing the unsequenceable: Expanded CGG-repeat alleles of the fragile X gene. Genome Res 23:121–128.

Mallawaarachchi AC, Hort Y, Cowley MJ, McCabe MJ, Minoche A, Dinger ME, Shine J, Furlong TJ. 2016. Whole-genome sequencing overcomes pseudogene homology to diagnose autosomal dominant polycystic kidney disease. Eur J Hum Genet 24:1584–1590.

Mardis ER. 2013. Next-Generation Sequencing Platforms. Annu Rev Anal Chem 6:287–303.

McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. 2010. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics 26:2069–2070.

Oliver GR, Hart SN, Klee EW. 2015. Bioinformatics for Clinical Next Generation Sequencing. Clin Chem 61:124–135.

Ozsolak F, Milos PM. 2011. Single-molecule direct RNA sequencing without cDNA synthesis. Wiley Interdiscip Rev RNA 2:565–570.

Qi X-P, Du Z-F, Ma J-M, Chen X-L, Zhang Q, Fei J, Wei X-M, Chen D, Ke H-P, Liu X-Z, Li F, Chen Z-G, et al. 2013. Genetic diagnosis of autosomal dominant polycystic kidney disease by targeted capture and next-generation sequencing: Utility and limitations. Gene 516:93–100.

Qiao W, Yang Y, Sebra R, Mendiratta G, Gaedigk A, Desnick RJ, Scott SA. 2016. Long-Read Single Molecule Real-Time Full Gene Sequencing of Cytochrome P450-2D6: Human Mutation. Hum Mutat 37:315–323.

Rossetti S, Consugar MB, Chapman AB, Torres VE, Guay-Woodford LM, Grantham JJ, Bennett WM, Meyers CM, Walker DL, Bae K, Zhang Q, Thompson PA, et al. 2007. Comprehensive Molecular Diagnostics in Autosomal Dominant Polycystic Kidney Disease. J Am Soc Nephrol 18:2143–2160.

Rossetti S, Hopp K, Sikkink RA, Sundsbak JL, Lee YK, Kubly V, Eckloff BW, Ward CJ, Winearls CG, Torres VE, Harris PC. 2012. Identification of Gene Mutations in Autosomal Dominant Polycystic Kidney Disease through Targeted Resequencing. J Am Soc Nephrol JASN 23:915–933.

Schirmer M, Ijaz UZ, D'Amore R, Hall N, Sloan WT, Quince C. 2015. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. Nucleic Acids Res 43:e37–e37.

Spithoven EM, Kramer A, Meijer E, Orskov B, Wanner C, Abad JM, Areste N, Alonso de la Torre R, Caskey F, Couchoud C, Finne P, Heaf J, et al. 2014. Renal replacement therapy for autosomal dominant polycystic kidney disease (ADPKD) in Europe: prevalence and survival--an analysis of data from the ERA-EDTA Registry. Nephrol Dial Transplant 29:iv15-iv25.

Steinberg KM, Schneider VA, Graves-Lindsay TA, Fulton RS, Agarwala R, Huddleston J, Shiryev SA, Morgulis A, Surti U, Warren WC, Church DM, Eichler EE, et al. 2014. Single haplotype assembly of the human genome from a hydatidiform mole. Genome Res 24:2066–2076.

Su Z, Ning B, Fang H, Hong H, Perkins R, Tong W, Shi L. 2011. Next-generation sequencing and its applications in molecular diagnostics. Expert Rev Mol Diagn 11:333–343.

Sun Y, Ruivenkamp CAL, Hoffer MJV, Vrijenhoek T, Kriek M, Asperen CJ van, Dunnen JT den, Santen GWE. 2015. Next-Generation Diagnostics: Gene Panel, Exome, or Whole Genome? Hum Mutat 36:648–655.

Tan AY, Michaeel A, Liu G, Elemento O, Blumenfeld J, Donahue S, Parker T, Levine D, Rennert H. 2014. Molecular diagnosis of autosomal dominant polycystic kidney disease using next-generation sequencing. J Mol Diagn 16:216–228.

Tan Y-C, Blumenfeld JD, Anghel R, Donahue S, Belenkaya R, Balina M, Parker T, Levine D, Leonard DGB, Rennert H. 2009. Novel method for genomic analysis of PKD1 and PKD2 mutations in autosomal dominant polycystic kidney disease. Hum Mutat 30:264–273.

Trujillano D, Bullich G, Ossowski S, Ballarín J, Torra R, Estivill X, Ars E. 2014. Diagnosis of autosomal dominant polycystic kidney disease using efficient PKD1 and PKD2 targeted next-generation sequencing. Mol Genet Genomic Med 2:412–421.

Vis JK, Vermaat M, Taschner PEM, Kok JN, Laros JFJ. 2015. An efficient algorithm for the extraction of HGVS variant descriptions from sequences. Bioinformatics 31:3751–3757.

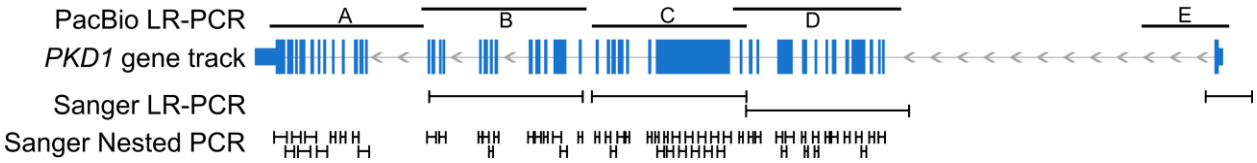
von K, Huber A. 2013. DNA methylation analysis. Swiss Med Wkly 143:w13799.

Wildeman M, Ophuizen E van, Dunnen JT den, Taschner PEM. 2008. Improving sequence variant descriptions in mutation databases and literature using the Mutalyzer sequence variation nomenclature checker. Hum Mutat 29:6–13.

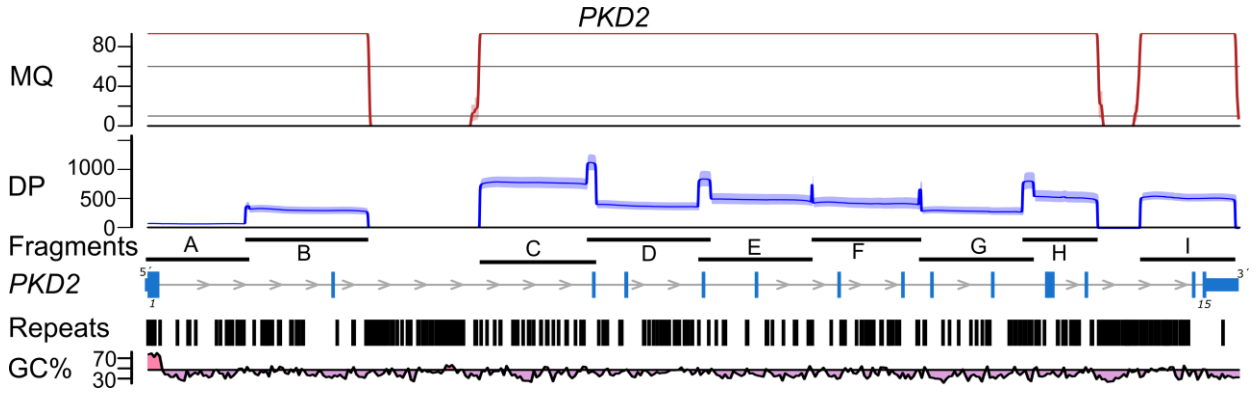
Willig LK, Petrikin JE, Smith LD, Saunders CJ, Thiffault I, Miller NA, Soden SE, Cakici JA, Herd SM, Twist G, others. 2015. Whole-genome sequencing for identification of Mendelian disorders in critically ill infants: a retrospective analysis of diagnostic and clinical findings. Lancet Respir Med 3:377–387.

Yang Y, Muzny DM, Reid JG, Bainbridge MN, Willis A, Ward PA, Braxton A, Beuten J, Xia F, Niu Z, Hardison M, Person R, et al. 2013. Clinical Whole-Exome Sequencing for the Diagnosis of Mendelian Disorders. N Engl J Med 369:1502–1511.

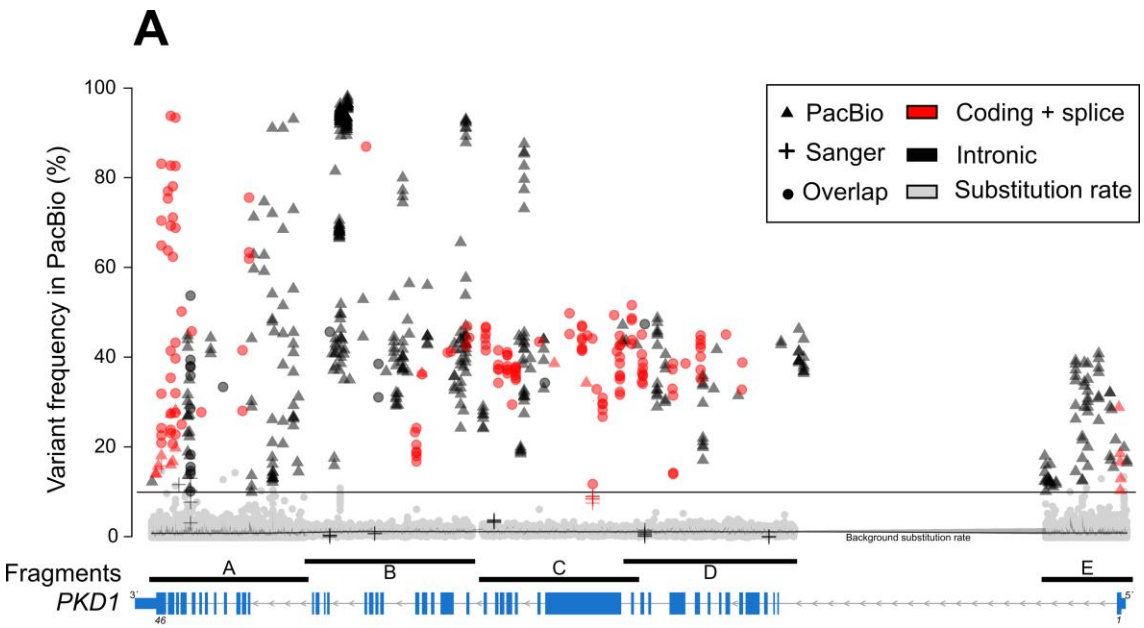
2.9. Annex I: supplementary material of Chapter 2

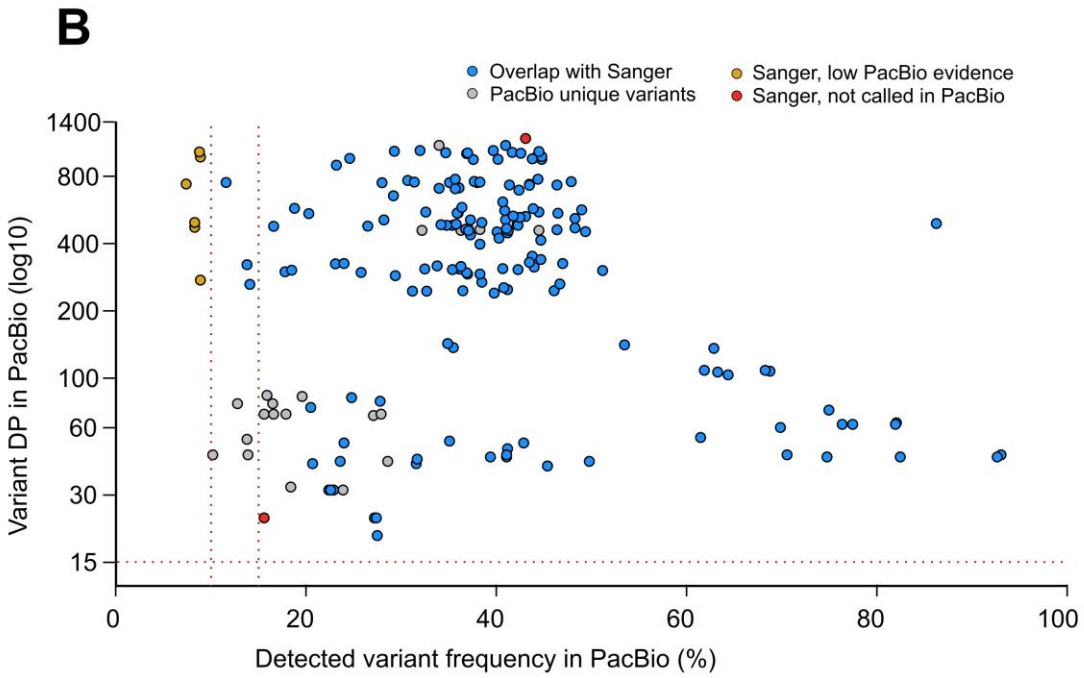


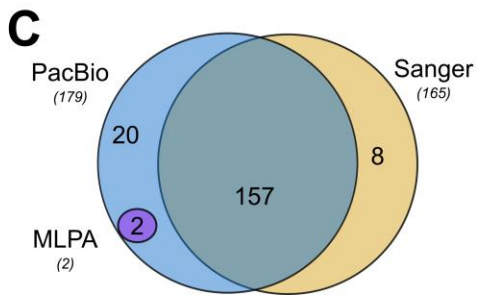
Supp. Figure S1: Enrichment of *PKD1* gene (NM_001009944.2) with 5 LR-PCR fragments used for PacBio, and 4 for Sanger sequencing followed by nested-PCR amplification for obtaining Sanger sequencing reads.



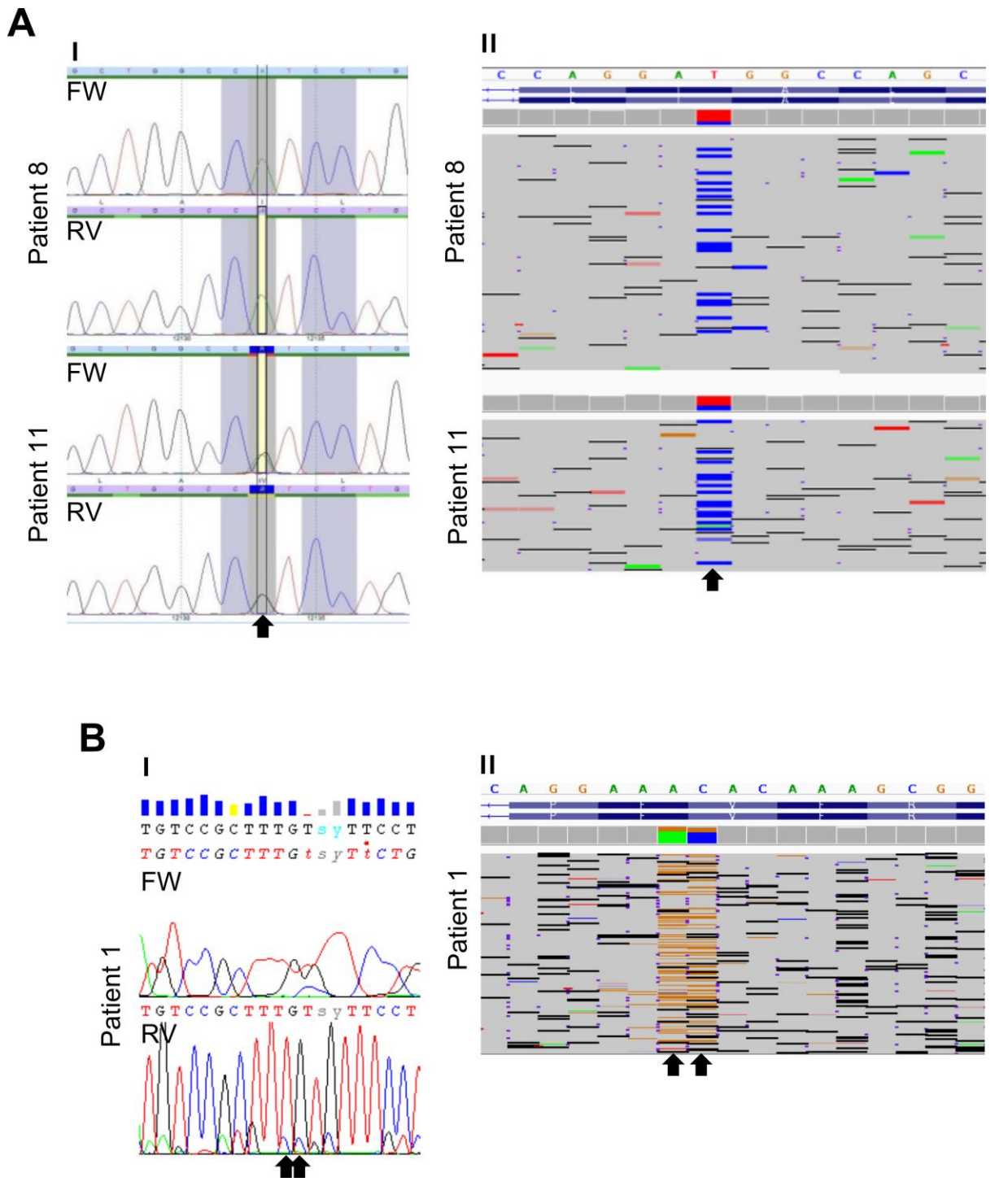
Supp. Figure S2: Mapping quality (MQ; in Phred quality scores; values >90 were scaled down for visualization purposes) and sequencing depth (DP; in number of reads) of uniquely aligned sequenced molecules to *PKD2* (NM_000297.3) for 9 LR-PCR fragments amplified. Mapping quality of alignments with even coverage distribution along the amplified fragments (Fragments), including regions with repetitive elements (Repeats) and high GC-content (GC%). Despite fragment A showing lower coverage compared to other fragments, it had sufficient coverage for variant calling within the exon regions, e.g. the first exon of *PKD2* with average coverage ≥71x (min. ≥43x; max. 111x) (Supp. Table S4).


A
Variant frequency in PacBio (%)
100
80
60
40
20
0
PacBio
Sanger
Overlap
Coding + splice
Intronic
Substitution rate
Background substitution rate
Fragments
A
B
C
D
E
PKD1
3'
5'
46
1

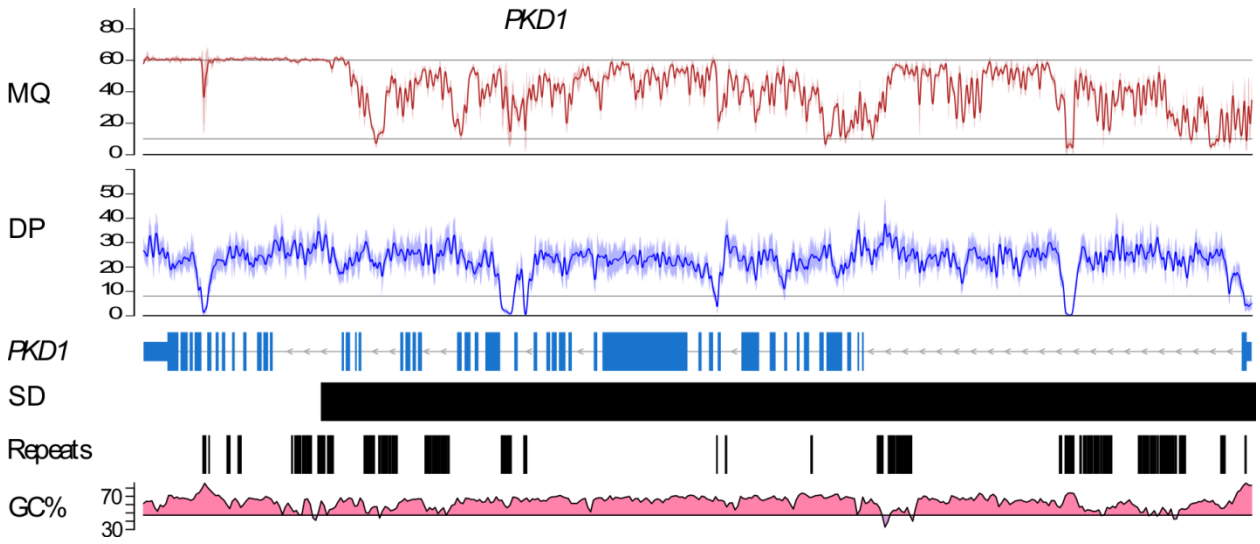

B
Overlap with Sanger
PacBio unique variants
Sanger, low PacBio evidence
Sanger, not called in PacBio
Variant DP in PacBio (log10)
1400
800
400
200
100
60
30
15
0
20
40
60
80
100
Detected variant frequency in PacBio (%)


C
PacBio
(179)
Sanger
(165)
20
157
8
MLPA
(2)
2

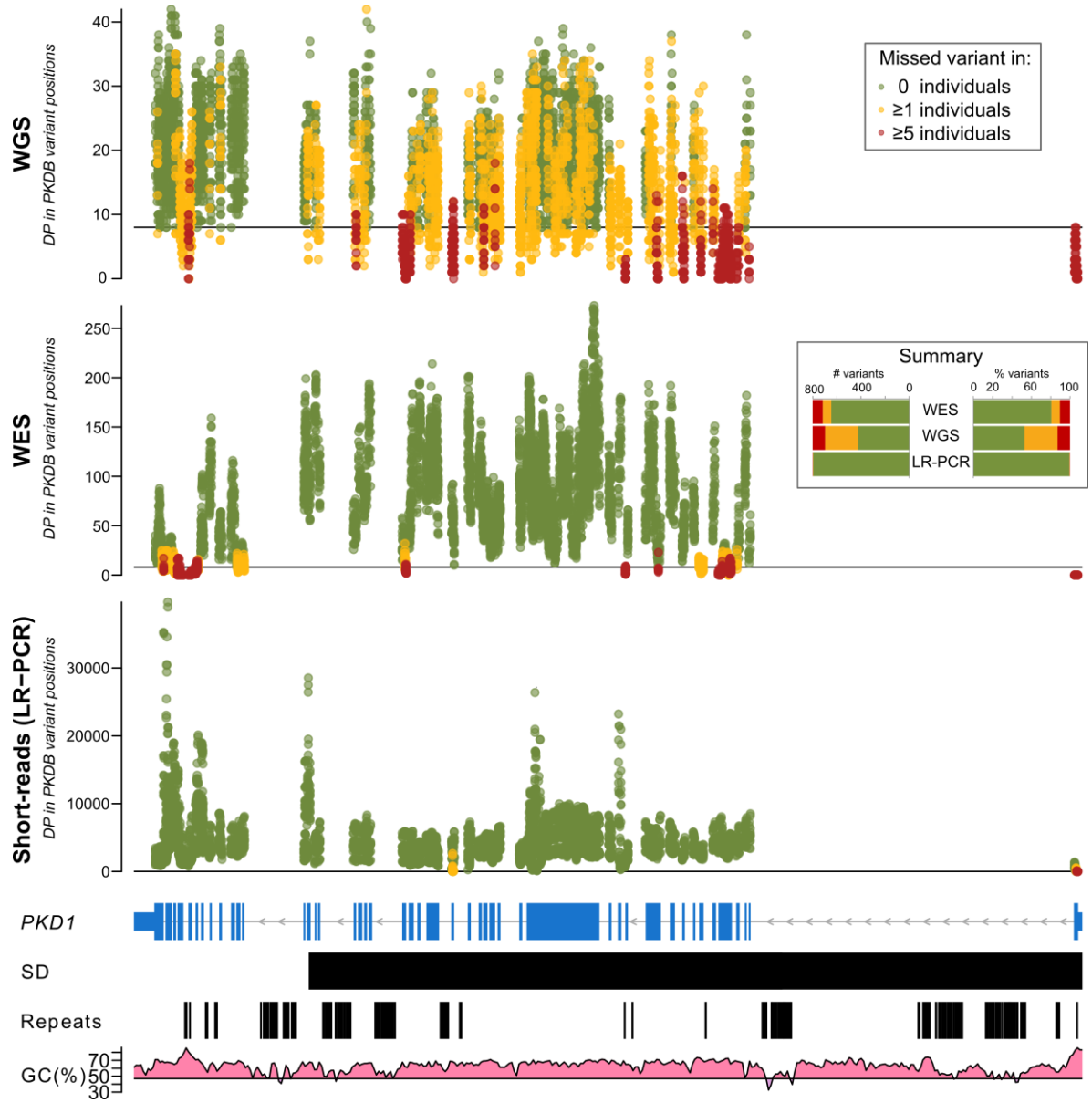
Supp. Figure S3: Comparison with current diagnostic assay. (*A.*) Small variants identified by PacBio were randomly distributed along the length of the amplified regions of *PKD1* gene locus (x axis; NM_001009944.2) showing no blind spots where no variants could not be detected. When compared to Sanger sequencing results, variants detected with both diagnostic approaches distributed over *PKD1* gene could be visualized (circles). The variant frequency (y axis) of variants solely detected by Sanger (crosses) or PacBio (triangles) which are majorly located in regions where no Sanger sequencing data was available, mostly introns (black). The average per-base mismatch rates (grey dots and average trend line) show the high sensitivity of our long-read sequencing method to detect substitutions. To highlight variants of high significance, coding or splice site variants (red) were colored to be visually identified from other intronic variants (black). (*B.*) The variant frequencies (x axis) with respect to the applied thresholds (dotted line) of all exonic and splice site variants detected either by Sanger or PacBio show the 2 Sanger variants that were not confirmed in PacBio (red) despite having sufficient sequencing depth (y axis) and variant frequency to be called. However, other 6 variants were potentially considered as PCR artifacts detected in Sanger sequencing despite high sequencing depth (yellow), which showed low variant frequency below the applied frequency thresholds (vertical dotted lines). Variants that passed the applied sequencing-based filtering thresholds and were detected by both diagnostic methods are shown as blue dots, whereas the variants that were detected by PacBio only are shown as grey dots. (*C.*) Overall, the comparison between our long-read sequencing approach with the standard ADPKD diagnostics assays show that from the 179 coding or splice site variants that were detected by PacBio, 2 were large deletions also identified by MLPA, and 157 were small variants also identified by Sanger (yellow). Only 8 small variants were not confirmed by PacBio from which 6 showed low sequencing data support (Supp. Figure S5B). The remaining 20 variants solely identified by PacBio comprehend 17 likely Sanger false negatives, and 3 low-confident calls of the allelic reconstruction of c.6657_6671del (Supp. Table S9).



Supp. Figure S4: Discordances identified between Sanger and PacBio sequencing are a clear example to show the sensitivity of direct long-read sequencing. (*A.*) The polymorphic substitution c.12133A>G correctly identified in patient 11 by Sanger (A_I) and PacBio (A_{II}) was clearly missed by Sanger sequencing in patient 8 (NM_001009944.2). (*B.*) Similarly, the polymorphism c.9195_9196delinsCC was not identified by Sanger sequencing due to unspecific forward signal (B_I), whereas in PacBio at least the substitution c.9195G>C was identified (B_{II}) (RefSeq NM_001009944.2).



Supp. Figure S5: Impact of *PKD1* (NM_001009944.2) inaccessible and unresolvable regions using NGS data. Mapping quality (MQ) and depth (DP) are affected by the presence of repetitive elements (Repeats), percentage of GC content (GC%), and segmental duplications (SD). Inaccessible and unresolvable regions, detected using a publicly available 9 WGS dataset (Sun et al. 2015), disrupt the MQ and DP where SDs from multiple *PKD1* pseudogenes interferes with the alignments of NGS short-reads. Dips in DP and MQ occur where repeats and high GC% overlap, even within gene coding regions (blue).



Supp. Figure S6: Loss of power to detect pathogenic variants with NGS approaches for *PKD1* gene (RefSeq NM_001009944.2). 12% of known pathogenic variants from the ADPKD database (PKDB) would be missed due to poor sequencing depth (red dots) regardless of the NGS sequencing strategy utilized for (WGS or WES) (exons 1, 5, 11, 12, 26, 35, 36, 41, 42 and 46). These poorly sequenced variants appear to be enriched in regions with segmental duplications (SD), repetitive elements (Repeats), and mainly with high percentage of GC-content (GC%) such is the case for exon 1. Short-read sequencing using LR-PCR enrichment of *PKD1* regions substantially improves the coverage of known pathogenic variants with only a 1.3% of missed variants due to low or poor coverage.

Supp. Table S1: Amplified LR-PCR fragments and sequencing pools for the SMRT library preparation of *PKD1* and *PKD2*.

Amplified fragments	Exons covered	Fragment size (Kb)	Sequencing pool*	Strand**	LR-PCR Primer sequence (5'-3')***
PKD1 A	35-46	7.5	B	FW	GAGCAGGCTCATGGGGCTTTGTAGGAGC
				RV	ACAGCCCGCTGTACCTGAGGACTCG
PKD1 B	22-34	8.1	C	FW	ATGTGAAGAGGTGCCTTGTGTGGT
				RV	TTAAAAACCCGCCCATAATTTCTCACTGC
PKD1 C	14-21	7.6	C	FW	GTTTCCCTGTCTGTTGGGAGGTAAC
				RV	CTGCGTTCACACAGGACAGAACG
PKD1 D	2-13	8	C	FW	CCGAGTAGCTGGAACTACAGTTACACACT
				RV	CACCCAGTTACCTCCCAACAGAC
PKD1 E	1	4.3	A	FW	GCGGAGCGTGAAAAATAGCTCGT
				RV	TACTGCTTTGCTTGACCAGCCTTAAAGA
PKD2 A	1	7.2	B	FW	GCAGGATTCTGTTGCTAGAAGTCAGTGC
				RV	CCTTTCTATCTAGCTTCTTTCCATCCCAGC
PKD2 B	2	7.9	C	FW	CCTGTAACTCCCACCATGGAATGGGC
				RV	AGGTAGGCTTGGAGGGTGCAACTGG
PKD2 C	3	7.5	B	FW	TACCCCTTAAAGATTTTCCTCACA
				RV	CTGTCGATACTCATGCATTGAAA
PKD2 D	3-5	8	C	FW	CTGTGTTGGGGCCTGTGCAGATCAGC
				RV	GGGGGACTTGGTGATGGAACATGTGGC
PKD2 E	5-6	7.4	B	FW	TTGGGACTACATTGACCTCACTAA
				RV	TTCATTCCTGTATCCCCAGTGC
PKD2 F	7-8	7.1	B	FW	GTTTTCTGAGCACCTACTATGTACTTGC
				RV	TAAACCTTGACAACAGTCACCCTCG
PKD2 G	9-10	7.4	B	FW	TGAACTCCAGGGCCTCACACTGTCC
				RV	GCGAACTTTAGACCTGACCTTGCTTTGC
PKD2 H	11-13	4.9	A	FW	AAATGTTGGGGCTGGACATGGTGGC
				RV	ATGCACAAGGAACATTCTTCAGGACG
PKD2 I	14-15	6.2	A	FW	CAGGTCTTTGTCTGCTAAGTCTGA
				RV	TTTGCAAGTGAAATGAAAAACAGT

Pooling of similar size fragments was performed to optimize loading and sequencing efficiency.
FW, forward strand; RV, reverse strand.
****Primers included a 5' M13-tail (TGTAAAACGACGGCCAGT for FW primers, and CAGGAAACAGCTATGACC for RV primers).***

Supp. Table S2: Per patient targeted PCR amplification and alignment statistics of number of reads and unique molecules, sequenced with SMRT-Seq.

Patient	# Subreads	# Unique molecules (reads)			Base coverage per exon (±20 bp)*		
		PKD1	*PKD2*	*PKD1 Pseudog.*	*>10x*	*>20x*	*>30x*
1	47,842	2,217	5,756	581	100%	100%	93.64%
2	31,889	1,932	3,611	427	100%	100%	99.95%
3	29,200	1,631	2,980	367	100%	100%	99.53%
4	16,929	1,008	1,719	158	100%	99.84%	84.63%
5	31,925	1,584	3,757	236	100%	100%	100%
6	23,344	1,405	2,413	159	100%	99.95%	89.32%
7	26,497	1,268	3,043	276	100%	100%	99.995%
8	36,737	2,208	4,469	202	100%	100%	100%
9	38,224	1,979	4,973	203	100%	99.15%	98.6%
10	44,782	2,667	4,976	684	100%	91.22%	84.12%
11	51,389	2,520	6,054	631	100%	100%	100%
12	42,132	2,357	5,043	412	100%	100%	100%
13	41,092	2,205	4,517	627	100%	100%	100%
14	46,520	2,530	5,010	570	100%	100%	100%
15	46,276	2,880	5,294	254	100%	100%	100%
16	48,624	3,958	5,543	562	100%	100%	100%
17	46,126	3,163	4,948	557	100%	100%	100%
18	44,009	2,671	5,485	334	100%	100%	99.99%
19	45,185	2,684	5,439	332	100%	100%	100%
Mean:	38,902	2,256	4,478	399	100%	99.48%	97.36%

Reads were filtered by read length >500 bp and read quality over 75% within SMRT Analysis. Only properly mapped sequences with mapping quality over >30 that were primary alignments and not PCR or optical duplicates were counted.

**Per exon percentages calculated using the total number of exon (±20 bp) bases from the targeted design (18,259 bp).*

Supp. Table S3: Total number of unique molecules (reads) sequenced for each amplicon.

Amplified fragments		Exons covered	*# Unique Molecules (reads)*			
			Total	Mean	Min.	Max.
PKD1	A	35-46	11,271	593	300	1,580
PKD1	B	22-34	14,439	760	396	1,814
PKD1	C	14-21	23,526	1,238	487	1,845
PKD1	D	2-13	24,281	1,278	519	1,830
PKD1	E	1	1,650	87	35	153
PKD2	A	1	7,982	420	221	737
PKD2	B	2	8,737	460	220	831
PKD2	C	3	24,132	1,270	480	1,748
PKD2	D	3-5	32,089	1,689	610	2,459
PKD2	E	5-6	25,433	1,339	478	2,079
PKD2	F	7-8	23,673	1,246	483	2,197
PKD2	G	9-10	24,187	1,273	436	2,196
PKD2	H	11-13	16,372	862	256	1,547
PKD2	I	14-15	12,007	632	240	886

Supp. Table S4: Base coverage statistics as average min, average mean, and average maximum, calculated from coverage stats retrieved for each patient and exon (±20bp).

Gene	***Exon***	**Base coverage**			***Gene***	***Exon***	**Base coverage**		
		Mean	*Min.*	*Max.*			*Mean*	*Min.*	*Max.*
PKD1	1	55	24	91	*PKD2*	1	71	43	111
PKD1	2	422	175	646	*PKD2*	2	303	111	617
PKD1	3	420	168	638	*PKD2*	3	1.097	388	1,509
PKD1	4	414	163	627	*PKD2*	4	384	122	672
PKD1	5	408	150	616	*PKD2*	5	833	270	1,249
PKD1	6	403	150	586	*PKD2*	6	477	121	866
PKD1	7	402	149	577	*PKD2*	7	452	190	984
PKD1	8	402	150	573	*PKD2*	8	436	167	1,138
PKD1	9	400	147	563	*PKD2*	9	295	49	572
PKD1	10	401	145	564	*PKD2*	10	272	47	632
PKD1	11	400	141	558	*PKD2*	11	514	137	766
PKD1	12	392	139	545	*PKD2*	12	508	134	762
PKD1	13	378	139	528	*PKD2*	13	491	130	748
PKD1	14	756	306	1,224	*PKD2*	14	507	183	714
PKD1	15	754	257	1,256	*PKD2*	15	510	188	711
PKD1	16	734	258	1,186					
PKD1	17	739	250	1,199					
PKD1	18	744	252	1,201					
PKD1	19	743	253	1,198					
PKD1	20	738	253	1,187					
PKD1	21	713	244	1,155					
PKD1	22	550	308	1,252					
PKD1	23	564	312	1,349					
PKD1	24	559	309	1,358					
PKD1	25	555	304	1,361					
PKD1	26	551	294	1,366					
PKD1	27	548	281	1,432					
PKD1	28	547	283	1,436					
PKD1	29	544	279	1,442					

PKD1	30	543	279	1,446
PKD1	31	544	275	1,525
PKD1	32	542	275	1,528
PKD1	33	532	269	1,521
PKD1	34	523	264	1,504
PKD1	35	59	24	138
PKD1	36	57	21	134
PKD1	37	56	22	133
PKD1	38	54	20	127
PKD1	39	53	21	125
PKD1	40	51	20	122
PKD1	41	51	20	118
PKD1	42	51	19	115
PKD1	43	51	19	113
PKD1	44	51	19	110
PKD1	45	51	21	110
PKD1	46	49	19	108

Supp. Table S5: Details of variants identified by the Sanger sequencing or PacBio long-read sequencing approach.

c. notation**	p.notation	Chr.	Pos.	# PB	# S	# O	PKDB*	dbSNP	Note
PKD1									
c.12897C>T	p.(=)	chr16	2,139,743	1	0	0			HC; D; Likely S FN
c.12890A>G	p.(Lys4297Arg)	chr16	2,139,750	1	0	0		rs758833703	HC; D; Likely S FN
c.12780T>C	p.(=)	chr16	2,139,860	1	0	0			HC; D; Likely S FN
c.12630T>C	p.(=)	chr16	2,140,010	8	8	7	LN	rs7203729 (C)	HC; Likely PB FN; Likely S FN
c.12409C>T	p.(=)	chr16	2,140,321	3	3	3	LN	rs79899502 (C)	HC
c.12276A>G	p.(=)	chr16	2,140,454	10	8	8	LN	rs3087632 (C)	HC; Likely S FN
c.12182C>T	p.(Ala4061Val)	chr16	2,140,548	1	1	1		rs372105572	HC; Likely S FP
c.12176C>T	p.(Ala4059Val)	chr16	2,140,554	4	3	3	LN	rs3209986 (C)	HC; Likely S FN
c.12133A>G	p.(Ile4045Val)	chr16	2,140,680	10	8	8	LN	rs10960 (C)	HC; Likely S FN
c.11916C>T	p.(=)	chr16	2,140,972	2	2	2	LN	rs77634115	HC
c.11682C>T	p.(=)	chr16	2,141,454	1	1	1	LN	rs567482892	HC
c.11554del	**p.(Leu3852TrpfsTer93)**	**chr16**	**2,141,581**	**1**	**1**	**1**	**DP**	**rs724159823**	**HC**
c.11412-3C>A	**p.?**	**chr16**	**2,141,910**	**1**	**1**	**1**			**HC**
c.10768C>T	p.(=)	chr16	2,143,865	2	2	2	LN	rs116114803	HC
c.10535C>T	p.(Ala3512Val)	chr16	2,144,176	3	3	3	LN	rs34197769 (C)	HC
c.9957C>T	p.(=)	chr16	2,149,738	1	1	1	LN	rs141101590	HC
c.9397+1G>A	**p.?**	**chr16**	**2,152,061**	**1**	**1**	**1**	**DP**		**HC**
c.9330T>C	p.(=)	chr16	2,152,129	6	6	6	LN	rs12926160 rs144582212 (C)	HC
c.9324del	**p.(Ile3109SerfsTer207)**	**chr16**	**2,152,134**	**1**	**1**	**1**		**rs780284643**	**HC**
c.9195_9196delinsCC	p.(Phe3066Leu)	chr16	2,152,387	5	5	5	LN	rs372874584	HC
c.9195G>C	p.(=)	chr16	2,152,388	1	0	0	LN	rs9935834 (C)	HC; D; Likely S FN
c.9187C>T	p.(Arg3063Cys)	chr16	2,152,396	1	1	1	LN	rs145906459	HC
c.9034_9039del	**p.(Thr3012_Ser3013del)**	**chr16**	**2,152,543**	**1**	**1**	**1**			**HC**
c.8859dup	**p.(Glu2954Ter)**	**chr16**	**2,152,903**	**1**	**1**	**1**			**HC**
c.8440G>A	p.(Gly2814Arg)	chr16	2,153,618	1	1	1	LN	rs149151043	HC
c.8293C>T	**p.(Arg2765Cys)**	**chr16**	**2,153,765**	**1**	**1**	**1**	**LH**	**rs144979397**	**HC**
c.8161+8G>A	p.?	chr16	2,154,491	1	1	1	LN	rs199569003	HC
c.8123C>T	p.(Thr2708Met)	chr16	2,154,537	1	1	1	LN	rs147350387	HC
c.8020C>T	p.(Pro2674Ser)	chr16	2,154,640	1	1	1	LN	rs144557371	HC
c.8017-2_8017-1del	**p.?**	**chr16**	**2,154,643**	**1**	**1**	**1**	**DP**		**HC**
c.7940C>T	**p.(Thr2647Met)**	**chr16**	**2,155,399**	**1**	**1**	**1**		**rs748496650**	**HC**
c.7913A>G	p.(His2638Arg)	chr16	2,155,426	5	5	5	LN	rs9936785 (C)	HC
c.7708T>C	p.(=)	chr16	2,156,021	5	5	5	LN	rs28575767 (C)	HC
c.7441C>T	p.(=)	chr16	2,156,447	7	7	7	LN	rs2003782 (C)	HC
c.7214G>T	**p.(Trp2405Leu)**	**chr16**	**2,156,674**	**1**	**1**	**1**			**HC**
c.7165T>C	p.(=)	chr16	2,156,850	7	7	7	LN	rs2457533 (C)	HC
c.6994_7000dup	**p.(Ala2332GlyfsTer90)**	**chr16**	**2,157,954**	**1**	**1**	**1**	**DP**		**HC**
c.6986G>A	p.(Arg2329Gln)	chr16	2,157,963	1	1	1	LN	rs575211353	HC
c.6670_6673del	p.(Pro2224TrpfsTer17)	chr16	2,158,494	1	0	0			Likely LAA Low Confidence Assembly
c.6657_6671del	**p.(Arg2220_Pro2224del)**	**chr16**	**2,158,496**	**1**	**1**	**1**			**HC; Complete PB LAA Confident Assembly**
c.6666_6667del	p.(Leu2223AlafsTer38)	chr16	2,158,500	1	0	0			Likely LAA Low Confidence Assembly
c.6656_6659del	p.(Pro2219ArgfsTer22)	chr16	2,158,508	1	0	0			Likely LAA Low Confidence Assembly
c.6488G>A	p.(Arg2163Gln)	chr16	2,158,680	1	0	0	LN	rs145217118	HC; D; Likely S FN
c.6223_6224insTT	**p.(Arg2075LeufsTer42)**	**chr16**	**2,158,944**	**0**	**1**	**0**			**Likely PB FN**
c.5763G>A	p.(=)	chr16	2,159,405	2	2	2	LN	rs2575313 (C)	HC

c.5172C>T	p.(=)	chr16	2,159,996	7	7	7	LN	rs9935526 (C)	HC
c.4968_4969delinsC	**p.(Arg1657GlyfsTer65)**	**chr16**	**2,160,198**	**1**	**1**	**1**			**HC**
c.4968T>C	p.(=)	chr16	2,160,200	1	0	0		rs777909326	HC; D; Likely S FN
c.4917C>T	p.(=)	chr16	2,160,251	1	1	1	LN	rs148852027	HC
c.4674G>A	p.(=)	chr16	2,160,494	1	1	1	LN	rs79884128 (C)	HC
c.4665A>C	p.(=)	chr16	2,160,503	1	7	1	LN	rs71385734 (C)	HC; D; Likely S FP
c.4475G>C	p.(Arg1492Pro)	chr16	2,160,693	1	1	1			HC
c.4248dup	**p.(Gly1417TrpfsTer14)**	**chr16**	**2,160,919**	**1**	**1**	**1**			**HC**
c.4195T>C	p.(Trp1399Arg)	chr16	2,160,973	5	5	5	LN	rs116092985 (C)	HC
c.3643C>G	p.(Leu1215Val)	chr16	2,161,525	1	1	1		rs144338515	HC
c.3513C>G	p.(=)	chr16	2,161,655	1	1	1	LN	rs143784787	HC
c.3375C>T	p.(=)	chr16	2,161,793	5	5	5	LN	rs74331768 (C)	HC
c.3372C>T	p.(=)	chr16	2,161,796	5	5	5	LN	rs75510884 (C)	HC
c.3275T>C	p.(Met1092Thr)	chr16	2,162,361	5	5	5	LN	rs2549677 (C)	HC
c.3111A>G	p.(=)	chr16	2,162,839	3	3	3	LN	rs2099534 (C)	HC
c.3100A>G	p.(Asn1034Asp)	chr16	2,162,850	1	1	1		rs369180760	HC
c.3063T>C	p.(=)	chr16	2,162,887	5	5	5	LN	rs2369068 (C)	HC
c.2730C>T	p.(=)	chr16	2,164,294	2	2	2	LN	rs35965348 (C)	HC
c.2700G>A	p.(=)	chr16	2,164,324	2	2	2	LN	rs35667726 (C)	HC
c.2694A>C	p.(=)	chr16	2,164,330	2	2	2	LN	rs142357713	HC
c.2681_2690del	**p.(Phe894Ter)**	**chr16**	**2,164,333**	**1**	**1**	**1**			**HC**
c.2269del	**p.(Gln757SerfsTer28)**	**chr16**	**2,164,754**	**1**	**1**	**1**		**rs775710328**	**HC**
c.2109C>T	p.(=)	chr16	2,164,915	1	1	1	LN	rs527655141	HC
c.1850-4A>G	p.?	chr16	2,165,630	7	7	7	LN	rs35929659 (C)	HC
c.1602C>T	p.(=)	chr16	2,166,838	1	1	1	LN	rs759092782	HC
c.1286G>T	**p.(Trp429Leu)**	**chr16**	**2,167,589**	**1**	**1**	**1**	**HLP**		**HC**
c.1261C>T	**p.(Arg421Cys)**	**chr16**	**2,167,614**	**1**	**1**	**1**			**HC**
c.182C>T	p.(Pro61Leu)	chr16	2,185,509	1	0	0	LN		HC; D; Likely S FN
c.129C>T	p.(=)	chr16	2,185,562	1	0	0			HC; D; Likely S FN
c.122C>T	p.(Pro41Leu)	chr16	2,185,569	2	0	0			HC; D; Likely S FN
c.96C>T	p.(=)	chr16	2,185,595	1	0	0			HC; D; Likely S FN
PKD2									
c.53C>T	p.(Pro18Leu)	chr4	88,928,938	1	0	0			D; Likely S FN
c.83G>C	p.(Arg28Pro)	chr4	88,928,968	11	9	9	LN	rs1805044 (C)	HC; Likely S FN
c.361G>T	p.(Gly121Cys)	chr4	88,929,246	1	1	1		rs371898195	HC
c.420G>A	p.(=)	chr4	88,929,305	5	5	5	LN	rs2728118 (C)	HC
c.568G>A	p.(Ala190Thr)	chr4	88,929,453	1	1	1	LN	rs117078377 (C)	HC
c.720C>G	p.(=)	chr4	88,957,382	0	2	0			D; Likely S FP
c.1459T>C	p.(Tyr487His)	chr4	88,967,933	1	1	1	LN	rs201328200	HC

** Bold=Pathogenic; DP=Definitely Pathogenic; HLP=Highly Likely Pathogenic; LH=Likely Hypomorphic; LN=Likely Neutral; C=Common; HC=High-Confident; FP=False Positive; FN=False Negative; D=Discordant; PB=PacBio; S=Sanger; O=Overlap.*

+ Deletion-insertions were manually inspected and HGVS description modified to avoid variant redundancies or undesired complex descriptions by clustering of neighbouring variants (e.g. c.9195_9196delinsCC and c.6994_7000dup). Individual substitutions c.9195G>C and/or c9196T>C from the polymorphic site c.9195_9196delinsCC were removed as redundant variants if c.9195_9196delinsCC was identified as well in the same patient. The pathogenic duplication c.6994_7000dup was originally detected as c.6982_6983delinsAGCGGCTG when reconstructing the PKD1 *allelic sequence and was split into two independent descriptions c.6994_7000dup, and c.6982del.*

*** The 177 coding or splice site variants listed in this table were either reported by Sanger sequencing or detected by our direct long-read sequencing approach in at least one patient within the cohort of 19 screened ADPKD patients. 157 variants are high-confidence variants identified on the same set of patients by both approaches. 8 Sanger variants were not confirmed by PacBio, from which 6 had low sequencing data support. The remaining 20 variants were solely detected by PacBio from which 3 are low-confident variants called by the reconstructions of allelic sequences of c.6657_6671del.*

*** RefSeq NM_001009944.2 or NM_000297.3 for PKD1 and PKD2, respectively.*

Supp. Table S6: Amplified LR-PCR primer sequences for Sanger sequencing.

Exons covered	***Fragment size (Kb)***	***Strand****	***LR-PCR Primer sequence (5'-3')****
1	2.2	FW	CGCAGCCTTACCATCCACCT
		RV	TCATCGCCCCTTCCTAAGCA
2-13	7.9	FW	CCGAGTAGCTGGAACTACAGTTACACACT
		RV	CACCCAGTTACCTCCCAACAGAC
14-21	7.6	FW	GTTTCCCTGTCTGTTGGGAGGTAAC
		RV	CTGCGTTCACACAGGACAGAACG
22-34	7.5	FW	CCGTGTAGAGAGGAGGGCGTGTGCAAGGA
		RV	TCGGCAAGGACCTGCTGGATCAG

**FW, forward strand; RV, reverse strand.*
***Primers included a 5' M13-tail (TGTAAAACGACGGCCAGT for FW primers, and CAGGAAACAGCTATGACC for RV primers).*

Supp. Table S7: Nested PCR primer sequences for *PKD1* Sanger sequencing.

Exons covered	**Size (bp)**	**Strand***	**Nested PCR Primer sequences (5'-3')****
1 *(frag. 1)*	256	FW	CCTGAGCTGCGGCCTCCG
		RV	CAGTTGACGCGGCAGGCG
1 *(frag. 2)*	216	FW	TGCGAGCCCCCCTGCCTC
		RV	AACCCGCCCACGCCCGCCCGTCC
2-3	340	FW	TAGGGGCTCTGGCCCTGAC
		RV	CCAGCCAGGACCCCACCCAAAG
4	266	FW	CATAGACCCTTCCCACCAG
		RV	CCTGGCTGGGAAGGACAGA
5 *(frag. 1)*	389	FW	TGGAGCCAGGAGGAGCAGAA
		RV	CAGAGGGACAGGCAGGCAAA
5 *(frag. 2)*	431	FW	AGCCCTCCAGTGCCTCCTTT
		RV	GCACGGCCGTCACGTGATAG
6	280	FW	ACCGTTGACACCCTCGTTCC
		RV	CTCTGCCCCAGTGCTTCAG
7	329	FW	CTGTGAGGGTGGGAGGATGG
		RV	CTAACCACAGCCAGCGTCTC
8	226	FW	GCGGCTCGGTCCCCAGTCT
		RV	GGAGGGCAGGTTGTAGAACGTG
9 *(frag. 1)*	295	FW	GGAGTCTGGGCTTCAGGCTG
		RV	CTGGGAACCACTCTGGTGGC
9 *(frag. 2)*	228	FW	GGAGTCTGGGCTTCAGGCTG
		RV	CACCCACCACCCAGAGTCCC
10 *(frag. 1)*	184	FW	GGCCTGTGGGCAAATCAGGG
		RV	TGGGGGTGGCAGGAGGCGTC
10 *(frag. 2)*	201	FW	AGGGGGACGCTGGTGCCCTG
		RV	GGGAACAGACCCAGGTCAGG
11 *(frag. 1)*	425	FW	GTCCACGGGCCATGACCG
		RV	CCAGCCACTGGGGAGACCAC
11 *(frag. 2)*	257	FW	GGCAGAGGTGGGCAATGG

		RV	AGCCGGGCACGAAGGTGGC
11 *(frag. 3)*	326	FW	GTGTCAGCGCCCGCTTTG
		RV	CTGTGTGAGCACCCTGTCTGC
12	241	FW	GTGTGTCCAGGAGGCGA
		RV	AGAGGTGAAGGTGGAGC
13	244	FW	CTGCCACCTGGGCTCACTG
		RV	TGCCCACCCCAAACCGGC
14	198	FW	CTCACTGCGTCCCACCGC
		RV	CTGAAAGGCAGTGGCCCC
15 *(frag. 1)*	382	FW	TGGGGAGCAGGTGGGGGTGC
		RV	AGACGCGCACATCCGCCTGGGCCG
15 *(frag. 2)*	363	FW	CGTGCGCCTGGAGGTCAAC
		RV	GGCTGCGTGGGGATGCAG
15 *(frag. 3)*	373	FW	CGTGCTGGTCTTCGTCCTGG
		RV	TGTAGCGGTAGGGGAACGG
15 *(frag. 4)*	380	FW	GTTTGTGCAGCTCGGGGAC
		RV	AAGCGTGGGTGACCTCCG
15 *(frag. 5)*	376	FW	CCCGCCAGCTACCTGTGG
		RV	GCGGAGCCCACCTCGTTC
15 *(frag. 6)*	378	FW	CTTCCGCTCCGTGGGCAC
		RV	GGAGGCGGCCACCATCAG
15 *(frag. 7)*	374	FW	AGCGCCTGGGCCGACTGCAC
		RV	AGCTGCCCCCAAAAGGGC
15 *(frag .8)*	363	FW	GAGCCCGGAGGCAGCTTC
		RV	GGGAGCACCTCGGGGTTG
15 *(frag. 9)*	377	FW	AGCTGTCACCTTCCGCCTG
		RV	GCACCTGGATCTCCAACAGCC
15 *(frag. 10)*	340	FW	GCTGGTCATCCTGTCGGGC
		RV	CACCAGGTTGGAGGCGTTC
15 *(frag. 11)*	567	FW	CCAGGGCCGAGCACTCCTAC
		RV	TGGGGTCGTAGGACTCGCTC
15 *(frag. 12)*	196	FW	CGCCTGGTGCCCATCATTG
		RV	GGACGGGTGAGGGGCATG
16	230	FW	AGGCCACGTCGCCCCTTG
		RV	GAGGCTGGGCTGTCCAAGG
17	203	FW	GAGGTAACCCCACTCCCACG
		RV	ATCCCCAGCCCGCCCACAC
18	353	FW	AGAGGGTTGCGCCCCCTC
		RV	ATCCCGCTGCTCCCCCCACGC AGG
19	286	FW	TCCCGTGATGCCGTGGGG
		RV	CAGGTGGCAGTCTCGGGG
20	260	FW	CCACCTGCTCACCACCCC
		RV	GCAGGGGTACAGGTCTTGGTC CC
21	226	FW	GCGCTGCTGACAGCTTGC
		RV	ATGCGGGGCAGGGTGAGC
22	220	FW	AGTGGGGCCAGGAGCGGG
		RV	GGGCGGGTGGCATGGGGC
23 *(frag. 1)*	349	FW	CCCTCCCTCTACCTCCCTGTC
		RV	CACTGAGGTTGGCCAGGGC
23 *(frag. 2)*	431	FW	GGGCCTGGCTGCCACTTC
		RV	AAGGCCAGGGGGCCGCGTG
24	222	FW	CAGGCGTGTGACCTGCGC
		RV	TGCCCTGCCCTGCCAGGCTG
25	313	FW	CTGGGCTCACGTCCGCTAC
		RV	GCTCCCAGGAGCACAGGGTC

26	289	FW	GAGAAGGCACAGCTTGCACG
		RV	AGAGCAGGGGAGGCCCTG
27	240	FW	GCAGACCGAGCCTCCCAC
		RV	AGGGGCAGAGCTTGGCAG
28	217	FW	TGCGAGCCTGACCTCCCTC
		RV	CCAACCTCCCACGGAGTGG
29	316	FW	TTGGGCAGGGTGGTCCTG
		RV	GGAAGGGCTGGGCAGGAAG
30	202	FW	CAGCCTCACCTGTGTGGCC
		RV	TCCATTCCCAGTACTCCCGG
31-32	338	FW	GAGCAGGTCTGAGCTGCCG
		RV	GCACCAGGGCTCGAGGTTTC
33-34	473	FW	GGTGGGCTGTGTGTGTGAC
		RV	CCCCTCCTCTGGCAATCC
35-36	576	FW	CAGGTTAACATGGGCTTGG
		RV	GAGGGGGTGGCTTCAGAG
37	341	FW	GGTAGGCTACAGGCCTCCAT
		RV	GAGGTGGGAGACAAGAGACG
38	304	FW	AAAGCCCTGCTGTCACTGTGG
		RV	TAGGGTCTGGCTGGACTAAAG
39	301	FW	GGGTCTCTGGTGGCCGCTCAC
		RV	ATGCCAGAGCTCCGCTAAAGG
40-41	565	FW	GCAGGAAACACTCCTGTTGG
		RV	GCTCCTGGCTGGTGACTG
42	608	FW	GAGCCCACCCTCACTCCTC
		RV	AACAGCAGCAGGCACACCT
43	660	FW	CTCTGCTCACCTCGGTACG
		RV	CGGACGAGAAATCTGTCTGC
44	361	FW	CGCCGCTTCACTAGCTTCGAC
		RV	AGTCCCAGGGCACAGCACCAA
45	486	FW	GGGAGGGCGTCTTAGCTC
		RV	CACAGGGGCTCAGTCAGTC
46	680	FW	CAAGGTCAAGGAGGTGGGTA
		RV	TTGACAGCGGCAGAAAGTAA

[*]FW, forward strand; RV, reverse strand.

[**]Primers included a 5' M13-tail (TGTAAAACGACGGCCAGT for FW primers, and CAGGAAACAGCTATGACC for RV primers).

Supp. Table S8: Nested PCR primer sequences for *PKD2* Sanger sequencing.

Exon covered	Size (bp)	Strand[*]	Oligo sequence (5'-3')[**]
1 *(frag. 1)*	421	FW	GAAAGGAACATGGCTCCTGA
		RV	ACCTCCTCCTCCTCCTCCTC
1 *(frag. 2)*	454	FW	CCCTTCTCCTCCGCTCTC
		RV	CGTTCTGGTTCGTGCATCT
02	320	FW	GTGCTTTATTTTCCCTTTTG
		RV	GGTGCATACACACTTCCTTT
03	330	FW	CTTTGTGAAGGCTGCTGGT
		RV	TCCTGTCGATACTCATGCATTG
04	435	FW	TTGGTTATGCAAACGATG
		RV	GAATGGTGGGAGTTCAGAG
05	414	FW	CCTCAAGTGTTCCACTGATT
		RV	GTAGCTAACTGCAGGCAAAG
06	401	FW	CTGGCTGTATTCATGTGTTG
		RV	AATGCTGAGGAGATCAAAGA

07	336	FW	GGTAAGTTTCATATTTCTAAAACACT
		RV	TTCCATGATTTTGTGGAACT
08	329	FW	CACACCATTTTGTTTATCCA
		RV	TTCTTGAGAAGCAGTGACAA
09	277	FW	TGCATCAACTAGTGGACATT
		RV	GAGAAGACAAGGATTTACGAAG
10	259	FW	TTCCAAATTATGTTTCTTCCTT
		RV	AAAATCTGGGTGAAACAATG
11	278	FW	AAAACAGATGCAAAAGGAGA
		RV	CCAGGAATTTATCTTTAGAAGC
12	266	FW	GAACTGGGTACAAGGAATGA
		RV	TTTGATACATCTGTGGTGTTG
13	322	FW	GTCCTTGGTGAGGCTTCT
		RV	CTGGTCTCATGTGGACTCTT
14	285	FW	AAAGACAATGACAAGCACTTT
		RV	TCATTAATAACACCATGCTCA
15	417	FW	ATTATTTGGTCCCTGGACTT
		RV	GTGCTTGTTACAGCAATTCA

[*]FW, forward strand; RV, reverse strand.

[**]Primers included a 5' M13-tail (TGTAAAACGACGGCCAGT for FW, and CAGGAAACAGCTATGACC for RV).

Supp. Table S9: Discordant variant calls between Sanger sequencing and PacBio long-read sequencing approach.

c. notation	p.notation	Chr	pos	# PB	# S	# O	PKDB*	dbSNP	Note
PKD1									
c.12897C>T	p.(=)	chr16	2,139,743	1	0	0			HC; D; Likely S FN
c.12890A>G	p.(Lys4297Arg)	chr16	2,139,750	1	0	0		rs758833703	HC; D; Likely S FN
c.12780T>C	p.(=)	chr16	2,139,860	1	0	0			HC; D; Likely S FN
c.12630T>C	p.(=)	chr16	2,140,010	8	8	7	LN	rs7203729 (C)	HC; Likely PB FN; Likely S FN
c.12276A>G	p.(=)	chr16	2,140,454	10	8	8	LN	rs3087632 (C)	HC; Likely S FN
c.12176C>T	p.(Ala4059Val)	chr16	2,140,554	4	3	3	LN	rs3209986 (C)	HC; Likely S FN
c.12133A>G	p.(Ile4045Val)	chr16	2,140,680	10	8	8	LN	rs10960 (C)	HC; Likely S FN
c.9195G>C	p.(=)	chr16	2,152,388	1	0	0	LN	rs9935834 (C)	HC; D; Likely S FN
c.6670_6673del	p.(Pro2224TrpfsTer17)	chr16	2,158,494	1	0	0			Likely LAA Low Confidence Assembly
c.6666_6667del	p.(Leu2223AlafsTer38)	chr16	2,158,500	1	0	0			Likely LAA Low Confidence Assembly
c.6656_6659del	p.(Pro2219ArgfsTer22)	chr16	2,158,508	1	0	0			Likely LAA Low Confidence Assembly
c.6488G>A	p.(Arg2163Gln)	chr16	2,158,680	1	0	0	LN	rs145217118	HC; D; Likely S FN
c.6223_6224insTT	**p.(Arg2075LeufsTer42)**	**chr16**	**2,158,944**	**0**	**1**	**0**			**Likely PB FN**
c.4968T>C	p.(=)	chr16	2,160,200	1	0	0		rs777909326	HC; D; Likely S FN
c.4665A>C	p.(=)	chr16	2,160,503	1	7	1	LN	rs71385734 (C)	HC; D
c.182C>T	p.(Pro61Leu)	chr16	2,185,509	1	0	0	LN		HC; D; Likely S FN
c.129C>T	p.(=)	chr16	2,185,562	1	0	0			HC; D; Likely S FN
c.122C>T	p.(Pro41Leu)	chr16	2,185,569	2	0	0			HC; D; Likely S FN
c.96C>T	p.(=)	chr16	2,185,595	1	0	0			HC; D; Likely S FN
PKD2									
c.53C>T	p.(Pro18Leu)	chr4	88,928,938	1	0	0			D; Likely S FN
c.83G>C	p.(Arg28Pro)	chr4	88,928,968	11	9	9	LN	rs1805044 (C)	HC; Likely S FN
c.720C>G	p.(=)	chr4	88,957,382	0	2	0			D; Likely S FP

** Bold=Pathogenic; LN=Likely Neutral; C=Common; HC=High-Confident; FP=False Positive; FN=False Negative;D=Discordant; PB=PacBio; S=Sanger; O=Overlap.*

*** RefSeq NM_001009944.2 or NM_000297.3 for PKD1 and PKD2, respectively.*

[+] *Discordances shown in this table are represented in Supp. Figure S2B as yellow dots for likely S FP variants,red dots for likely PB FN variants, and grey dots for Likely S FN.*

Supp. Table S10: Variants detected in homopolymer stretches using long-read sequencing approach.

c. notation**	p.notation	Chr	pos	# PB	# S	# O	PKDB*	dbSNP	Note
PKD1									
c.12530del	p.(Pro4177HisfsTer21)	chr16	2,140,109	3	0	0		rs767438361	HS; Likely PB Artifact
c.12518del	p.(Pro4173ArgfsTer25)	chr16	2,140,121	5	0	0		rs778397103	HS; Likely PB Artifact
c.12445-3del	p.?	chr16	2,140,197	5	0	0		rs770813339	HS; Likely PB Artifact
c.12139-5del	p.?	chr16	2,140,595	2	0	0	LN	rs146430229	HS; Likely PB Artifact
c.12085del	p.(Val4029SerfsTer10)	chr16	2,140,727	1	0	0		rs781278135	HS; Likely PB Artifact
c.11713-5del	p.?	chr16	2,141,179	8	0	0			HS; Likely PB Artifact
c.11240del	p.(Pro3747HisfsTer79)	chr16	2,142,509	6	0	0			HS; Likely PB Artifact
c.10948del	p.(His3650ThrfsTer34)	chr16	2,143,612	1	0	0			HS; Likely PB Artifact
c.10822-8del	p.?	chr16	2,143,746	12	0	0	LN	rs373684171 rs9924796 (C)	HS; Likely PB Artifact
c.10745del	p.(Pro3582ArgfsTer3)	chr16	2,143,887	11	0	0	DP	CD076868	HS; Likely PB Artifact
c.9518del	p.(Pro3173ArgfsTer143)	chr16	2,150,446	1	0	0		rs772608027	HS; Likely PB Artifact
c.9176del	p.(Pro3059GlnfsTer15)	chr16	2,152,406	10	0	0		rs759773922	HS; Likely PB Artifact
c.9097del	p.(Leu3033TrpfsTer41)	chr16	2,152,485	1	0	0			HS; Likely PB Artifact
c.8586del	p.(Ile2863SerfsTer12)	chr16	2,153,471	1	0	0	DP		HS; Likely PB Artifact
c.8427del	p.(Glu2810ArgfsTer65)	chr16	2,153,630	1	0	0		rs746703342	HS; Likely PB Artifact
c.8019del	p.(Ser2675AlafsTer10)	chr16	2,154,640	1	0	0			HS; Likely PB Artifact
c.7864-3del	p.?	chr16	2,155,477	1	0	0		rs756848270	HS; Likely PB Artifact
c.7622del	p.(Pro2541ArgfsTer79)	chr16	2,156,172	1	0	0		rs538031465	HS; Likely PB Artifact
c.7401del	p.(Asn2468ThrfsTer152)	chr16	2,156,486	7	0	0		rs745812853	HS; Likely PB Artifact
c.6759del	p.(Glu2254SerfsTer60)	chr16	2,158,408	18	0	0			HS; Likely PB Artifact
c.6469del	p.(Leu2157CysfsTer4)	chr16	2,158,698	1	0	0			HS; Likely PB Artifact
c.5824del	p.(Arg1942AlafsTer7)	chr16	2,159,343	18	0	0		rs780100275	HS; Likely PB Artifact
c.5784del	p.(Glu1929ArgfsTer20)	chr16	2,159,383	1	0	0			HS; Likely PB Artifact
c.4485del	p.(Ala1496ProfsTer38)	chr16	2,160,682	1	0	0		rs578064441	HS; Likely PB Artifact
c.4269del	p.(Thr1424ProfsTer8)	chr16	2,160,898	1	0	0			HS; Likely PB Artifact
c.4220del	p.(Pro1407ArgfsTer25)	chr16	2,160,947	3	0	0		rs140412120	HS; Likely PB Artifact
c.4069del	p.(Leu1357TrpfsTer9)	chr16	2,161,098	16	0	0	DP	CD085910	HS; Likely PB Artifact
c.3684del	p.(Val1229TrpfsTer44)	chr16	2,161,483	6	0	0	DP	rs781384791	HS; Likely PB Artifact
c.3240del	p.(Ser1081ArgfsTer23)	chr16	2,162,395	2	0	0		rs777145613	HS; Likely PB Artifact
c.3099del	p.(Asn1034MetfsTer4)	chr16	2,162,850	12	0	0	DP	rs372461622	HS; Likely PB Artifact
c.2854-5del	p.?	chr16	2,163,297	19	0	0		rs114846412	HS; Likely PB Artifact
c.2823del	p.(Glu942ArgfsTer9)	chr16	2,164,200	1	0	0		rs767322474	HS; Likely PB Artifact
c.2494del	p.(Arg832AlafsTer66)	chr16	2,164,529	19	0	0	DP		HS; Likely PB Artifact
c.2222del	p.(Pro741ArgfsTer44)	chr16	2,164,801	1	0	0		rs779605081	HS; Likely PB Artifact
c.2085del	p.(Ala696ArgfsTer89)	chr16	2,165,390	19	0	0		rs760496344	HS; Likely PB Artifact
c.2037del	p.(Tyr680MetfsTer105)	chr16	2,165,438	13	0	0			HS; Likely PB Artifact
c.1987del	p.(Gln663ArgfsTer122)	chr16	2,165,488	11	0	0	DP		HS; Likely PB Artifact
c.1914del	p.(Ala639ArgfsTer146)	chr16	2,165,561	11	0	0	DP	rs777208671	HS; Likely PB Artifact
c.1386-3del	p.?	chr16	2,167,056	11	0	0			HS; Likely PB Artifact
c.1221del	p.(Ser408ArgfsTer57)	chr16	2,167,653	4	0	0		rs768893401	HS; Likely PB Artifact
c.771del	p.(Thr258ProfsTer32)	chr16	2,168,221	11	0	0			HS; Likely PB Artifact
c.755del	p.(Pro252ArgfsTer38)	chr16	2,168,237	19	0	0			HS; Likely PB Artifact
c.198del	p.(Ala67ArgfsTer6)	chr16	2,185,492	1	0	0			HS; Likely PB Artifact
c.108del	p.(Cys37AlafsTer36)	chr16	2,185,582	14	0	0			HS; Likely PB Artifact
c.78del	p.(Arg28AlafsTer45)	chr16	2,185,612	13	0	0			HS; Likely PB Artifact
PKD2									
c.128del	p.(Pro43ArgfsTer74)	chr4	88,929,012	2	0	0	DP	CD982885	HS; Likely PB Artifact
c.203del	p.(Pro68ArgfsTer49)	chr4	88,929,087	19	0	0	DP	rs751221093	HS; Likely PB Artifact
c.538del	p.(Leu180TrpfsTer53)	chr4	88,929,422	6	0	0			HS; Likely PB Artifact
c.783del	p.(Val262CysfsTer55)	chr4	88,957,444	6	0	0		rs766343471	HS; Likely PB Artifact
c.1003del	p.(Gln335ArgfsTer3)	chr4	88,959,561	18	0	0			HS; Likely PB Artifact

** DP=Definitely Pathogenic; LN=Likely Neutral; C=Common; HS=Homopolymer Stretch; PB=PacBio; O=Overlap.*
*** RefSeq NM_001009944.2 or NM_000297.3 for PKD1 and PKD2, respectively.*

Chapter 3

3. Genetic drivers of Acute Kidney Injury: transcription factor signature with potential tissue damage protective effect

Daniel M. Borràs[1,2,3], Maria D. Sanchez-Niño[4], Joost P. Schanstra[2,3], Bart Janssen[1], Alberto Ortiz[4]

[1]*GenomeScan B.V, Plesmanlaan 1d, 2333BZ Leiden, The Netherlands.*

[2]*Institut National de la Santé et de la Recherche Médicale (INSERM), Institut of Cardiovascular and Metabolic Disease, Toulouse, France.*

[3]*Université Toulouse III Paul-Sabatier, Toulouse, France.*

[4]*Instituto Investigación Sanitaria - Fundación Jiménez Díaz - Universidad Autónoma de Madrid and Fundación Renal Íñigo Álvarez de Toledo - Instituto Reina Sofía de Investigación Nefrológica, Madrid, Spain.*

3.1. Abstract

Acute kidney injury (AKI) can occur in hospitalized and critically ill patients, particularly with sepsis, and admitted into the intensive care unit with up to 20% to 60% prevalence. This syndrome of kidney damage or failure occurs within a brief lapse of time, few hours to few days, with high mortality within those critically ill patients (≈50%). There is currently no therapy that reduces the severity or accelerates recovery from AKI. Previous research suggested that *Gdf15* gene was associated with a protective mechanism against kidney tubular damage. We used *Gdf15* knock-out mice to investigate these mechanisms, using an unbiased RNA sequencing approach for experimental nephrotoxic AKI (folic acid nephropathy in mice). This allowed us to identify differentially expressed genes that associated with biological processes and pathway in an AKI generic response, as well as in a *Gdf15*-driven AKI response. Our results show that *Gdf15* has a primary role in the regulation of processes involving proliferation, inflammation, necrosis, fibrosis, and apoptosis among other processes regulated by these pathways. In addition, we could identify the transcription factor signatures that are potentially modulating the generic AKI and *Gdf15*-specific AKI responses, that could be used for to the discovery of new therapeutic targets for AKI.

3.2. Introduction

Acute kidney injury (AKI) is a syndrome of kidney damage or failure that occurs within a brief lapse of time, from few hours to few days. The reduction of kidney function causes accumulation of products in blood that should have been filtered by normal kidney function (Bellomo et al., 2012). AKI episodes are a frequent event in clinical patients, such as surgery or critically ill patients, as well as in patients treated with certain drugs, including antibiotics, immunosuppressant drugs, cancer chemotherapy, non-steroidal anti-inflammatory drugs and blood pressure lowering drugs, among others (Bellomo et al., 2012). For these critically ill patients, mortality during the AKI episode hovers around 50% (Bellomo et al., 2012). AKI increases the risk of death for at least a year after its occurrence and contributes to progression of chronic kidney disease (CKD) (Breit et al., 2012). However, there is no therapy that reduces the severity or accelerates recovery from AKI. Therapeutic intervention is limited to replacement of renal function by hemodialysis or renal transplantation (Vallon, 2016). Clearly, an improved understanding of the molecular drivers of AKI is required.

Growth differentiation factor 15 (*Gdf15*), also known as macrophage inhibitory cytokine-1 (*MIC-1*), is a member of the transforming growth factor β (*TGF-β*) superfamily. It has been

identified in an acute liver injury and regeneration study (Hsiao et al., 2000). *Gdf15* was expressed in a wide range of tissue and cell types. Expression was increased in response to stress and injury such as tissue hypoxia, oxidative stress, inflammation, and acute injury (Hsiao et al., 2000; Breit et al., 2012; Hellemons et al., 2012; Mazagova et al., 2013; Adela and Banerjee, 2015; Sándor et al., 2015; Yatsuga et al., 2015; Li et al., 2016). In CKD patients undergoing hemodialysis within the first 3 years of starting the treatment, circulating *Gdf15* was increased and independently associated to an increased mortality risk (Breit et al., 2012). Furthermore, it was identified as a predictive marker for transition between albuminuria stages in type 2 diabetes mellitus patients (Hellemons et al., 2012). However, the *Gdf15* gene does not appear to be a net contributor to tissue injury. Rather, its expression may represent an adaptive, tissue protective response. According to Adela and Banerjee., 2015, *Gdf15* protects endothelial cells from cellular injury, apoptosis and cell death in cardiovascular and diabetes diseases by inhibiting Jnk (c-Jun N-terminal kinase) or activating Smad signaling pathways among other signaling pathways (Adela and Banerjee, 2015). Furthermore, *Gdf15* gene knock-out (KO) mice displayed more severe tubular and interstitial damage, but no effect on the glomerular sclerosis rate, suggesting a protective effect of *Gdf15* in a preclinical model of diabetes-associated CKD (Mazagova et al., 2013). *Gdf15* seems to protect from kidney damage reducing inflammatory cell recruitment and enhancing tubular epithelial cell proliferation (Mazagova et al., 2013). However, the molecular pathways modulated by *Gdf15* during kidney injury, and its potential role in AKI have not been fully characterized (Breit et al., 2012; Mazagova et al., 2013; Vallon, 2016).

We have now addressed the molecular mechanisms contributing to AKI and their modulation by *Gdf15* in mice, using an unbiased RNA sequencing approach to evaluate the gene expression profiles of wild type (WT) and *Gdf15* knock-out (GDF15-KO) mice kidneys.

3.3. Methods

3.3.1. Mice strains, Gdf15 null and AKI induction

A total of 28 mice were studied, including WT, GDF15-KO, and transgenic mice overexpressing Gdf15 (GDF15-Tg) as positive control. Knock-out mice (GDF15-KO) were developed using a target construct transfection into R1 embryonic stem cells and were kindly donated by Prof. Se-Jin Lee, Johns Hopkins University (Hsiao et al., 2000). Transgenic GDF15-Tg that ubiquitously express human GDF15 protein under the control of CAG, using traditional pronuclear microinjection on the C57/BL6 background, were kindly donated by

Prof Thomas Eiling (NIH-NIEHS) (Baek et al., 2006). Mice were maintained at IIS-Fundación Jiménez Díaz (FJD), Madrid. AKI was induced in WT and GDF15-KO mice by a single injection of folic acid. Mice received a single intraperitoneal injection of folic acid (Sigma) 250 mg/kg in 0.3 mol/L sodium bicarbonate or vehicle and were sacrificed 24 h after injection (Ortiz et al., 2017). In total, 5 groups of mice were sacrificed at 24h of induction that correspond to WT controls without AKI induction (WT-Control, n=6), WT with AKI induction (WT-AKI, n=6), GDF15-KO controls without AKI induction (GDF15-KO-Control, n=6), GDF15-KO with AKI induction (GDF15-KO-AKI, n=7), and GDF15-Tg without AKI induction (n=3). Studies were conducted in accordance with the NIH Guide for the Care and Use of Laboratory Animals (Guide for the Care and Use of Laboratory Animals: Eighth Edition, 2011).

3.3.2. Tissue sampling and RNA isolation

The kidneys were perfused in situ with cold saline before removal. Half of the kidney from each mouse was fixed in buffered formalin, embedded in paraffin and used for immunohistochemistry and the other half was snap-frozen in liquid nitrogen for RNA and protein studies and stored at -80 °C for preservation of the RNA (Ortiz et al., 2017). Total RNA isolation was performed using (Pure Link RNA mini kit, Ambion) following the manufacturer's standard guidelines.

3.3.3. Messenger RNA library preparation and sequencing

Total RNA samples were sequenced at GenomeScan B.V., and processed under the ISO/IEC 17025 accredited messenger-RNA (mRNA) sequencing pipeline. The enrichment of mRNA species was performed by poly-A capture with oligo-dT magnetic beads (NEBNext Ultra Directional RNA Library Prep Kit for Illumina). Sample libraries were prepared following manufacturer's protocol (NEB #E7420S/L). Then, the quality and yield of the RNA was measured before and after the sample preparation with a Fragment Analyzer (Thermo Fisher) (Table 1). Resulting mRNA library size was checked for consistency with the expected size distribution (300 to 500 bps) and concentration. Sequencing was performed using an Illumina HiSeq 2500 platform for a 125 paired-end (PE) library.

Table 1: Quality control checks of RNA samples. Before (Sample QC), and after the process of mRNA sequencing library preparation (Library QC).

	Sample QC			***Library QC***				
Sample	*Conc. (ng/µl)*	*RQN*	*28S/18S*	*ng/µL*	*nmol/L*	*Avg. Size*	*Reads (#)*	*Quality (%>=Q30)*
WT Control 1	793.689	9.7	1.3	3.486	12.23	469	55,927,139	89.81
WT Control 2	650.2	9.6	1.3	3.6948	13.556	448	52,310,573	90.32
WT Control 3	241.5	9.8	1.4	4.1204	14.742	460	51,165,664	90.48
WT Control 4	359.7	9.8	1.3	3.7006	12.274	496	49,532,713	89.44
WT Control 5	334.4	9.6	1.3	2.2532	8.212	451	57,573,731	90.44
WT Control 6	417.6	9.5	1.3	3.6536	12.896	466	63,430,234	89.91
WT AKI 1	846.7	9	1.4	4.2538	15.478	452	66,589,846	89.85
WT AKI 2	1,180.9	8.7	1.7	3.8178	13.706	458	43,832,642	89.67
WT AKI 3	1,240.3	9.3	1.5	3.6762	13.12	461	58,700,358	89.87
WT AKI 4	201.9	9.1	1.3	2.4916	8.942	458	41,122,213	89.84
WT AKI 5	695.4	9.5	1.3	2.699	9.65	460	43,823,846	90.24
WT AKI 6	965.9	8.3	0.8	1.0934	3.78	476	42,771,320	89.87
GDF15 KO Control 1	290.8	9.1	1.2	2.0286	7.22	462	40,444,400	89.95
GDF15 KO Control 2	313.1	8.3	1.1	1.9946	6.97	471	50,653,012	89.73
GDF15 KO Control 3	336.4	8.8	1.1	2.1004	7.532	459	50,140,826	90.05
GDF15 KO Control 4	490.6	8.7	1.3	2.9698	10.93	447	53,706,306	90.41
GDF15 KO Control 5	201.2	8	1.2	1.4302	5.082	463	57,499,527	89.85
GDF15 KO Control 6	155.3	9	1.3	2.9312	10.476	460	43,979,500	89.76
GDF15 KO AKI 1	2,363.3	9.4	1.6	2.526	9.014	461	43,064,788	90.05
GDF15 KO AKI 2	1,203.7	9.6	1.6	2.6226	9.508	454	47,802,649	89.87
GDF15 KO AKI 3	1,793.2	9.2	1.7	1.937	6.818	467	40,799,842	89.33
GDF15 KO AKI 4	1,018.6	9.4	1.5	2.2208	7.876	464	49,150,240	89.9
GDF15 KO AKI 5	2,797.4	9.2	1.7	1.6674	5.83	471	40,630,787	89.78
GDF15 KO AKI 6	2,399.3	9.5	1.6	2.272	8.16	458	43,275,128	89.88
GDF15 KO AKI 7	667.4	9.8	1.5	2.9904	11.114	443	51,500,862	89.97
GDF15 Tg Control 1	1,283.3	6.1	1.1	0.5224	1.686	510	40,877,158	89.55
GDF15 Tg Control 2	1,124.5	4.9	0.3	1.081	3.778	471	44,055,503	89.45
GDF15 Tg Control 3	1,232.4	5.3	1.3	2.64	9.292	468	47,609,579	89.73

3.3.4. Data analysis

Sequence read clipping and filtering was performed using fast-A quality files (fastQ) for PE sequenced reads (Table 1), pre-processed to remove sequencing adapters and low quality base calls. Using Trimmomatic v.0.36 we performed the adapter clipping, minimum read length filtering (>20 bps), and average base quality trimming of Q >15 Phred on a 5 consecutive base sliding window.

Alignment and feature quantification for mouse reference sequence GRCm38. On average, >49M PE reads per sample (Table 1) were aligned to the mouse reference sequence version GRCm38 patch 4. Reference files were downloaded from GENCODE version M9 with the

corresponding feature annotation tables as GFF/GTF files. Alignment was performed using the STAR v.2.4.2a aligner with a two-step alignment approach to incorporate known and novel junction indexes. The first alignment was performed against an indexed reference containing known splice sites, and the second against a newly re-built indexed reference including the novel splice sites inferred from the first alignment step. Sorted alignments were stored as sequence alignment map (B/SAM) files for further processing. Only high quality alignments (Phred score > 30) (Table 1) for Gencode M9 annotated features were counted using HTSeq-counts v.0.6.1p1. Gene expression counts were then uploaded to R statistical computing software v.3.3.1 (R), and further processed with the RNA-seq analysis package DESeq2 v.1.10.1 (DESeq) to perform library size normalization, and differential expression comparisons.

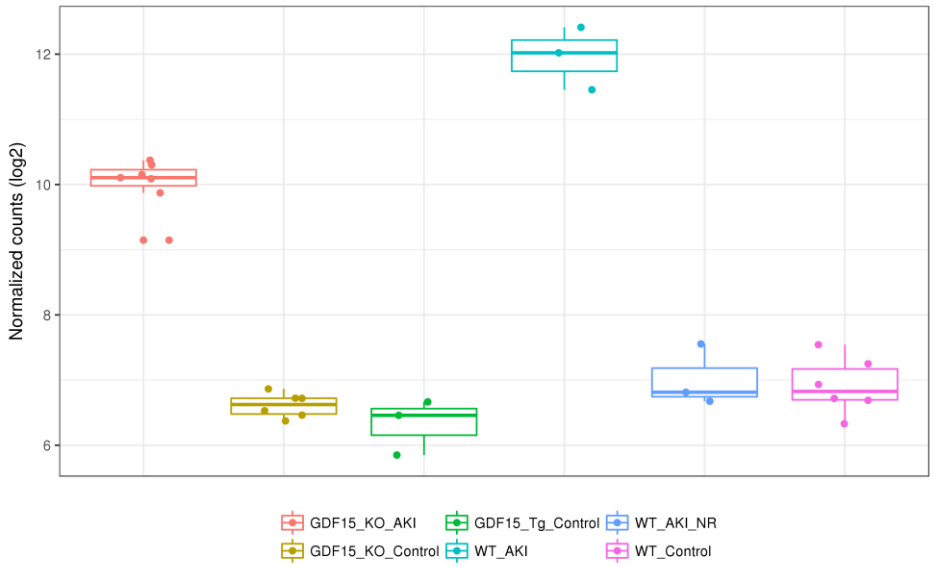


Figure 1: Expression of Gdf15 in log2 transformed normalized read counts. For all 5 groups of mouse samples plus 3 non-responder WT mice (WT_AKI_NR).

Effectiveness of gene deletion procedure to create GDF15-KO *mice* was qualitatively evaluated by assessing the expression levels and local alignments of *Gdf15* for all 28 samples. Normalized *Gdf15* expression in gene counts (Figure 1) was contrasted with *Gdf15* alignments (Supp. Figure S1) to confirm GDF15-KO and validate the functionality of *Gdf15* transcripts.

Confirmation of data-driven sample groups and clusters defined in the experimental design was assessed by data-driven group similarities within and between groups. Stratification or

batch effects of samples and groups was assessed by principal component analysis (PCA) (Figure 2). Then, unsupervised hierarchical clusters were generated from the normalized gene expression data, and hierarchical cluster probabilities were calculated with a bootstrap approach using pvclust v.2.0 package for R (Supp. Figure S2). The profile of the top 100 expressed genes, as well as sample distance matrices, were also evaluated as a measure to confirm the major trends of the observed sample clusters and grouping (Supp. Figure S3; Supp. Figure S4). Possible sample outliers, such as WT-AKI-NR, deviating from the original grouping were then separated for further downstream analysis as shown in the PCA plot

(Figure 2).

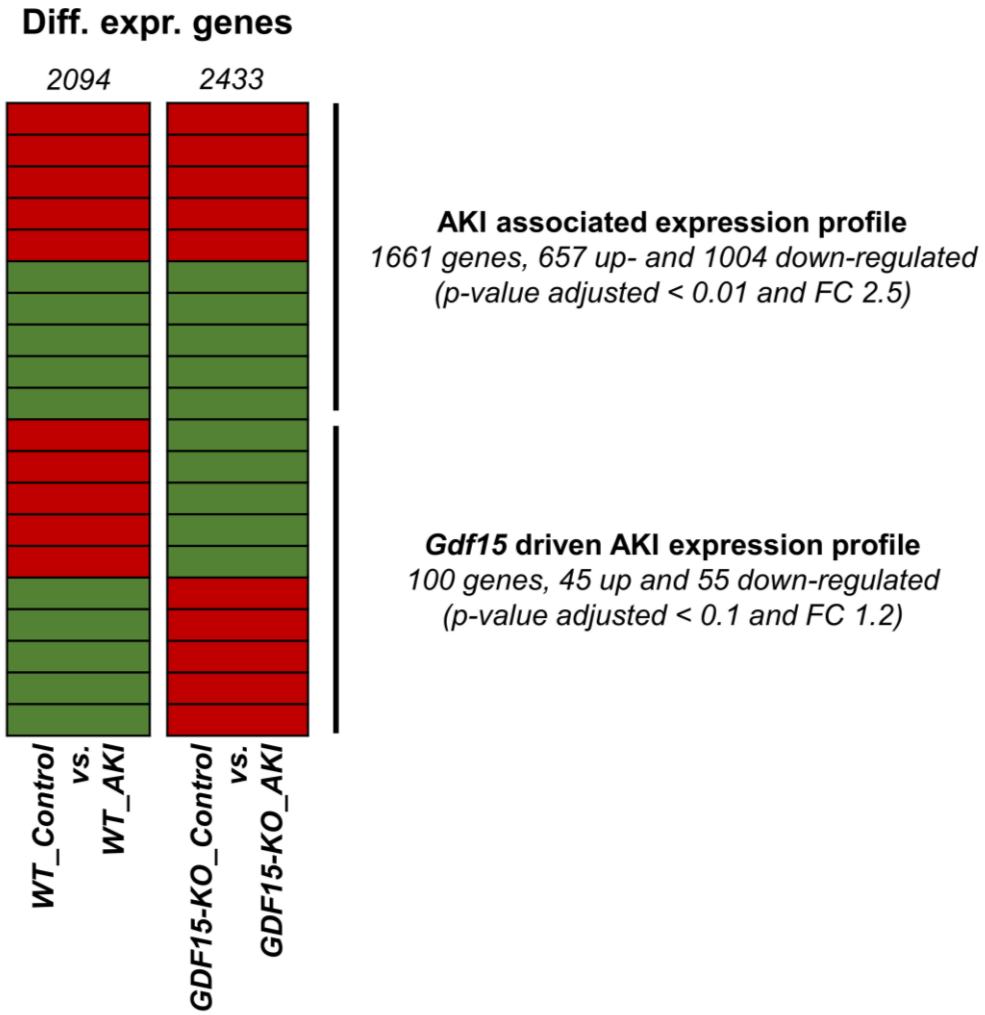


Figure 2: Scheme of the differential expression analysis for up- (red) and down-regulated (green) gene expression profiles associated with AKI and AKI potentially driven by *Gdf15*.

Differential expression comparison of sample groups was performed to identify the genetic differences associated with each annotated factor. Statistical comparisons were performed using DESeq package to identify the most significant features of sample group phenotypes. The intersection of differentially expressed genes in *Wt_Control vs. WT_AKI* and GDF15-KO-*Control vs.* GDF15-KO-*AKI* was extracted to assess the genetically driving differences between AKI treated samples and their respective controls. Bonferroni adjusted p-value < 0.01 (p-adj), and fold change 2.5 (FC) with the same numerical sign in both comparisons were considered as AKI associated differentially expressed genes (Figure 2). Opposite trends, were considered as an AKI response associated with *Gdf15* (p-adj < 0.1; FC 1.2 with opposite numerical sign) (Figure 2).

Enrichment analysis and functional annotation (gene ontology) was performed by evaluating the lists of differentially expressed genes as a network statistical visualization using ClueGO package v.2.2.6 (ClueGO) for Cytoskape v.3.4.0 (Cytoskape). For each comparison, groups of genes defined as up- or down-regulated were used as input categories in ClueGO to perform the enrichment and clustering of gene ontology (GO) biological processes. Visualization parameters were manually inspected to obtain an unsaturated and clearer network visualization.

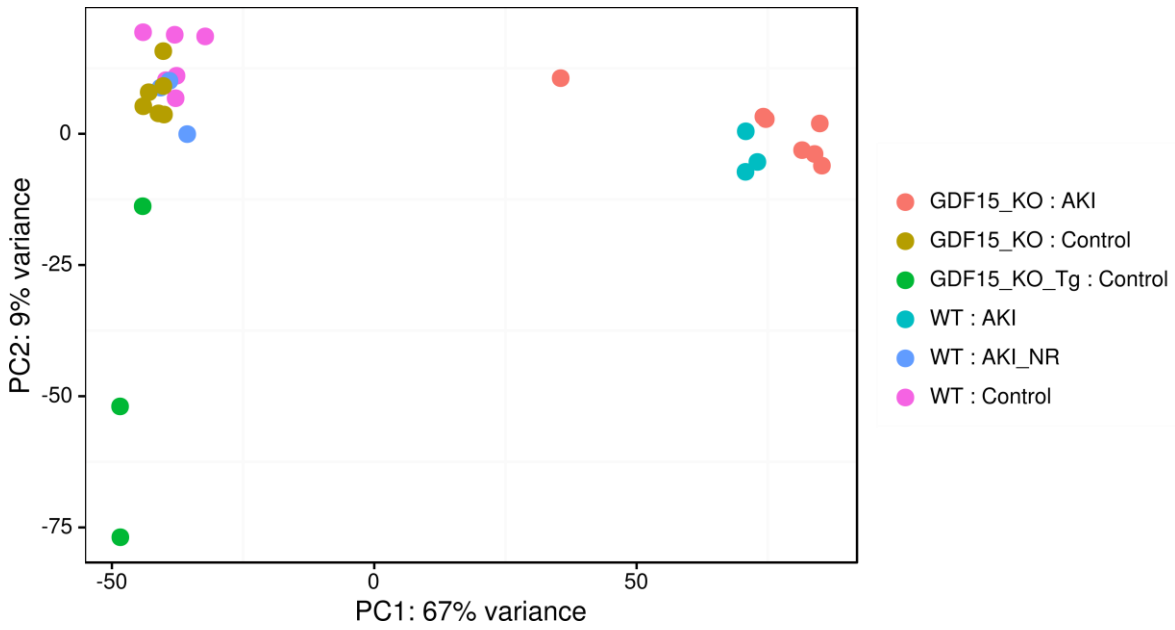


Figure 3: Principal component analysis of all 28 mRNA samples. PC1 (x axis) that accounts for 67% of the variance observed in the dataset is shown to discriminate between AKI and Control samples. PC2 (y axis) accounts for only 9% of the variance within the dataset and seems to be linked with GDF15-Tg samples.

Enrichment of renal and AKI disease-terms associated genes was performed using a reduction of identified differentially expressed genes by selecting those previously associated or linked with a disease term. Disease-term associations were obtained using a STRING database querying for a: genes associated with kidney disease or CKD, ischemia, nephrosclerosis, renal fibrosis, and kidney failure; and b: STRING genes associated with renal injury extracted from a PubMed query such as *[renal[All Fields] OR ("kidney"[MeSH Terms] OR "kidney"[All Fields])) AND ("wounds and injuries"[MeSH Terms] OR ("wounds"[All Fields] AND "injuries"[All Fields]) OR "wounds and injuries"[All Fields] OR "injury"[All Fields]]*. The list of disease genes included 2001 genes associated with renal injury, 1900 kidney disease, 1603 ischemia, 1035 kidney failure, 715 chronic kidney disease, 359 renal fibrosis, and 112 nephrosclerosis for a total of 3645 unique disease-terms associated genes.

Transcription factors driving the expression profiles were identified by comparing the differentially expressed genes lists with the transcription factors described for mouse model in the animal transcription factor database for (AnimalTFDB) (Zhang et al., 2012), or the Riken Transcription factor database (TFdb) (http://genome.gsc.riken.jp/TFdb/).

Evaluation of pathway involved in the AKI gene expression response was assessed using Pathvisio v.3.2.2 for the pathway statistical analysis. Biological pathway files were obtained from Pathvisio wiki pathways collection for *Mus musculus*, publicly available online (Kutmon et al., 2016). Statistics were based on an hypergeometric distribution test coupled to 1000 time permutation test of Z-scores for p-value calculation (Kutmon et al., 2015). Input data provided included all detected genes (> 1 read) and gene section for positive hits of p-adj < 0.1 and FC 1.2.

3.4. Results

3.4.1. High quality messenger-RNA sequencing identified wild type mice not responsive to folic acid induction

Deep-sequencing mRNA high quality libraries were obtained for a total of 28 kidney RNA samples (>49M PE reads on average). Alignments against the GRCm38 reference sequence, showed an exome representation of average 76.8% (Table 2). Observed GC-content of 47.4% was higher than the average genomic GC-content of 41.7% (), but matched the average exonic GC-content of 47.5% (Supp. Figure S5). Messenger RNAs expression values were

normalized to account for differences related to sequencing and the library preparation for each independent library.

The assessment of the high-quality expression profiles, combined with available sequence alignments, could identify 3 WT-AKI samples (WT-AKI-1, 4, and 5) that presented different expression profiles more correlated with non-AKI WT profiles. These samples were marked as outliers by re-naming them as non-responders (WT-AKI-NR) (Figure 1;Figure 3; Supp. Figure S1). Expression of top 100 genes within these three samples showed similar levels as the WT-Controls (Supp. Figure S3), and was confirmed by the hierarchical bootstrap analysis (Supp. Figure S2), as well as the sample distance vectors (Supp. Figure S4).

Table 2: RNA-sequencing of poly-A captured RNA. Basic read and alignment statistics for the 28 kidney mice samples.

Sample	*Aligned*	*Aligned pairs*	*Insert size*	*Phred*	*Aligned to exons (%)*		*3′/5′*	*GC%*
WT Control 1	114,706,762	56,837,513	368	22.1	78,211,324	80.54%	1.15	47.0
WT Control 2	105,028,232	52,028,891	337	24.3	68,366,112	77.49%	1.18	47.3
WT Control 3	104,061,772	51,572,275	345	23.1	70,136,718	79.65%	1.17	47.3
WT Control 4	103,924,539	51,448,537	388	24.4	68,438,647	78.89%	1.2	46.6
WT Control 5	116,569,558	57,755,378	338	22.0	75,679,482	77.35%	1.19	47.3
WT Control 6	129,245,563	64,035,007	340	23.7	84,145,297	77.50%	1.22	47.0
WT AKI 1	132,256,229	65,327,172	303	25.7	79,893,457	73.11%	1.21	48.1
WT AKI 2	88,034,650	43,446,603	331	25.6	55,550,948	75.57%	1.17	48.3
WT AKI 3	121,080,566	59,744,258	347	23.0	76,178,387	76.82%	1.16	47.7
WT AKI 4	83,158,844	41,168,751	341	26.9	52,497,798	75.87%	1.27	46.9
WT AKI 5	88,499,263	43,832,027	346	23.9	57,572,442	77.50%	1.2	47.2
WT AKI 6	86,548,478	42,739,683	342	23.2	55,024,939	77.07%	1.16	47.5
GDF15 KO Control 1	81,828,667	40,523,760	346	24.0	52,704,895	77.47%	1.24	46.9
GDF15 KO Control 2	103,670,604	51,323,280	355	23.8	67,287,247	78.44%	1.25	46.8
GDF15 KO Control 3	101,109,634	50,074,753	329	25.2	63,937,069	76.03%	1.25	46.9
GDF15 KO Control 4	107,641,180	53,333,009	326	23.6	68,792,743	76.12%	1.22	47.2
GDF15 KO Control 5	115,295,416	57,081,040	345	24.3	72,929,399	75.81%	1.23	47.2
GDF15 KO Control 6	89,583,323	44,349,691	353	23.8	57,812,739	76.96%	1.22	47.1
GDF15 KO AKI 1	87,793,188	43,373,268	362	23.3	57,144,431	77.77%	1.1	48.0
GDF15 KO AKI 2	96,434,008	47,607,202	361	22.6	62,633,298	77.15%	1.15	48.2
GDF15 KO AKI 3	83,631,609	41,239,226	364	23.4	53,331,754	77.46%	1.16	48.0
GDF15 KO AKI 4	100,577,250	49,671,665	372	22.7	65,623,653	77.90%	1.13	47.9
GDF15 KO AKI 5	83,588,255	41,219,940	354	22.4	52,164,956	76.91%	1.16	48.0
GDF15 KO AKI 6	87,202,229	43,128,322	362	22.4	56,930,660	77.61%	1.15	47.7
GDF15 KO AKI 7	106,248,349	52,409,037	354	22.2	67,812,747	78.07%	1.16	48.0
GDF15 Tg Control 1	86,469,404	42,633,873	348	32.6	53,167,273	76.83%	1.27	46.4
GDF15 Tg Control 2	90,144,863	44,123,702	333	38.9	49,546,706	68.13%	1.23	46.3
GDF15 Tg Control 3	96,297,277	47,635,345	345	26.7	59,691,384	75.29%	1.22	46.9
Average	99,665,347	49,273,686	348	24.6	63,685,947	76.83%	1.19	47.3

3.4.2. Underlying AKI-driving gene expression mechanisms, independent from Gdf15 expression

Overall gene expression response to AKI, independent from Gdf15, showed that 1661 genes are differentially expressed (p-adj < 0.01), and regulated in the same manner up- or down-regulated (FC 2.5) in both comparisons ("Wt_Control vs. WT_AKI" and "GDF15-KO-Control vs. GDF15-KO-AKI"), from which 1422 are protein coding (Figure 2). The first comparison WT_Control vs. WT_AKI yielded 2094 differentially expressed genes (1773 protein coding) and had a 79.3% of gene overlap with the GDF15-KO comparison (Figure 2). The GDF15-KO comparison (GDF15-KO vs. GDF15-KO-AKI) yielded 2433 differentially expressed genes (1992 protein coding) and showed a 68.2% of gene overlap with the WT comparison (Figure 2).

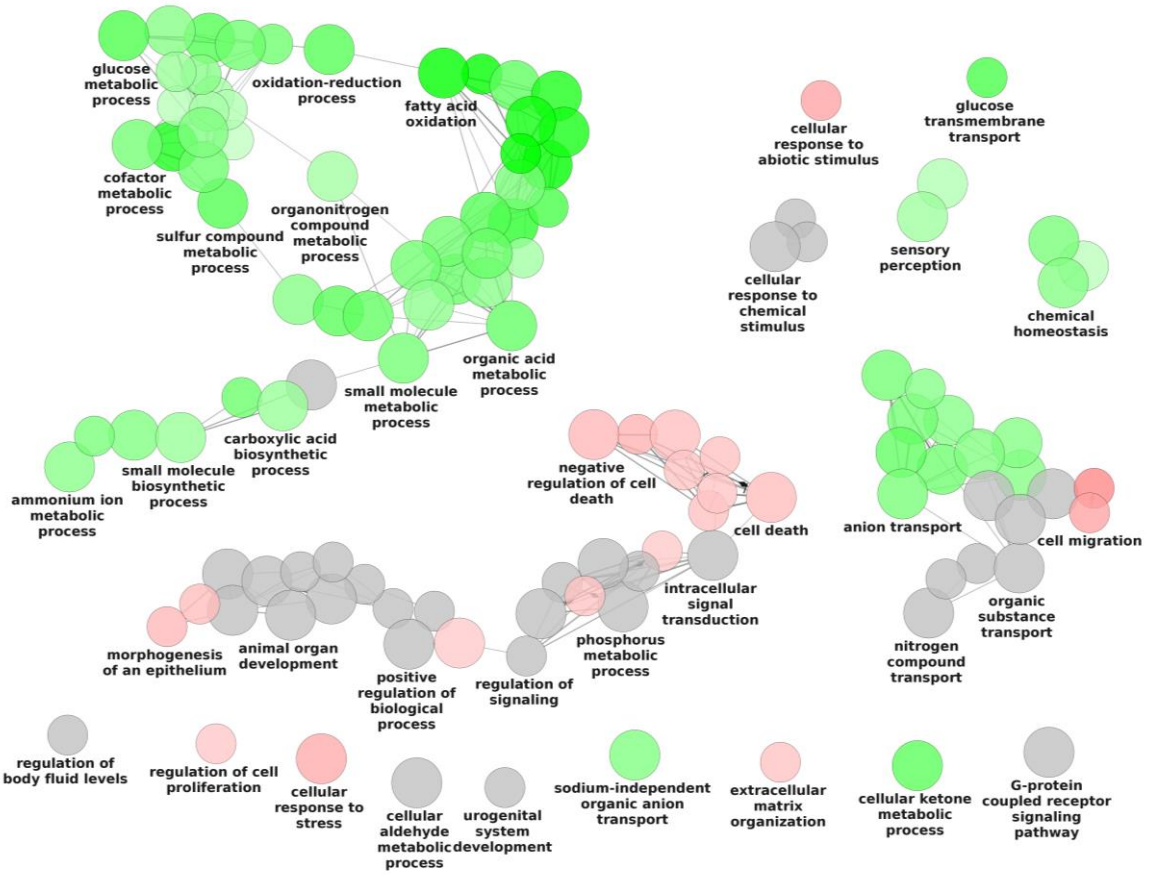


Figure 4: Enriched Gene Ontology Biological Processes lead by differentially expressed genes that are independent of Gdf15 protective mechanisms. Up- and down-regulated genes in both WT and GDF15-KO mice during AKI were considered to be Gdf15-independent. Circles represent enriched terms that are shown to be up-regulated (red), down-regulated (green), or no clear up- or down-regulated tendency (grey) during AKI. Terms have been clustered by term or process similarities and only the description of the major process that defines the cluster is shown (text). GO parent-child relationships between terms are shown as lines.

Gene ontology biological processes driven by differentially expressed genes in response to AKI show the underlying processes in AKI that are independent of *Gdf15* expression (Figure 3). Up- and down-regulated genes in both WT and GDF15-KO mice comparisons during an AKI episode produce a separated clustering of enriched biological process: up-regulated genes were driving cell regulation processes mainly involved in cell death and regulation of cell death, regulation of cellular processes, cell migration, proliferation and response to stress among others (Figure 3; Supp. Figure S6); whereas down-regulated genes were regulating metabolic processes, such as small molecule and organonitrogen compound metabolic processes, anion transport, cofactor metabolic process, chemical homeostasis, and oxidation-reduction processes among other metabolic processes as well as transport processes (Figure 3; Supp. Figure S7).

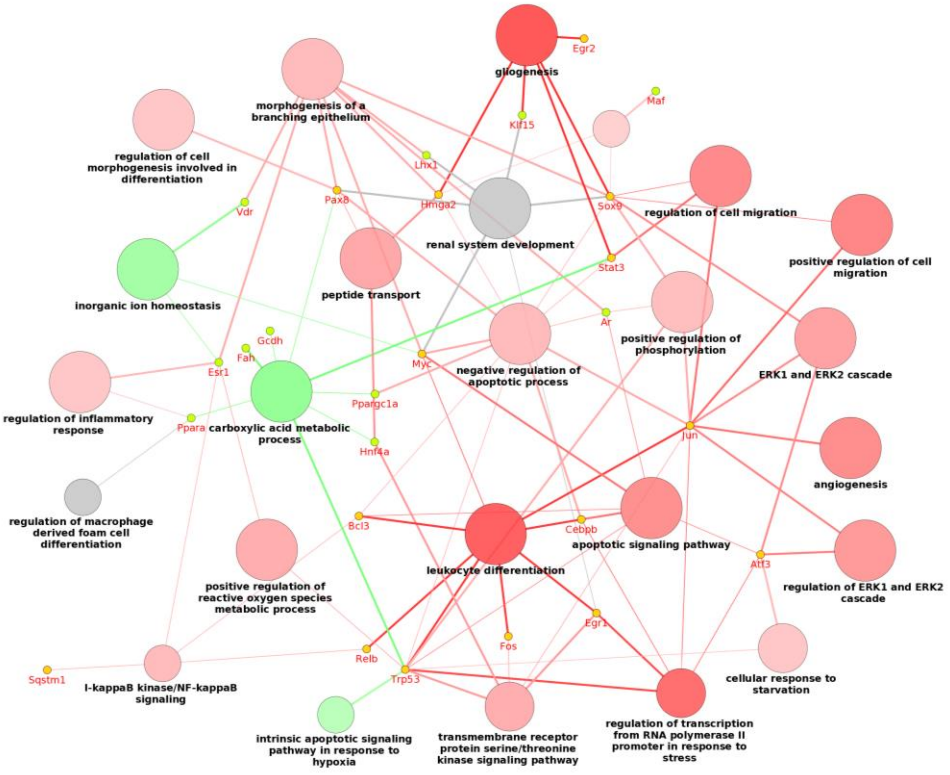


Figure 5: Enriched Gene Ontology Biological Processes lead by differentially expressed and disease associate genes with transcription factor activity that are independent of Gdf15 protective mechanisms. Circles represent enriched terms that shown to be up-regulated (red), down-regulated (green), or no clear up- or down-regulated tendency (grey) in WT-AKI mice. Terms have been clustered by term or process similarities and only the description of the major process that defines the cluster is shown (text). GO parent-child relationships between terms are shown as lines.

Evaluation of disease-term associated genes in AKI expression profile identified 415 protein-coding genes involved in renal disorders that were highly differentially expressed in the dataset (96.4% p-adj <0.001). In addition, we identified 1 long intergenic non-coding RNA (*Pvt1*), and 3 *"non-coding*" processed transcripts or pseudogenes (*Sngh1*, *Mpcp1-ps*, and *Pdxk-ps*) also associated with disease-terms. The identified disease-associated genes in AKI show to be up-regulating processes such as cell regulation, morphogenesis, and migration processes such as positive regulation of cell migration, morphogenesis of epithelial tube, negative regulation of apoptosis, and regulation of inflammatory response among other immune system, transport and morphogenic processes and responses (Supp. Figure S8). The down-regulated genes were enriched in terms associated with kidney development, an/ion transport, inorganic ion homeostasis, water transport, apoptotic signaling in response to hypoxia, and macrophage derived cell differentiation (Supp. Figure S9).

Table 3: Disease-associated transcription factors driving an AKI response in both comparisons "WT_Control vs. WT_AKI" and "GDF15-KO-Control vs. GDF15-KO-AKI".

	GDF15-KO changes in AKI			***WT changes in AKI***			***Genomic information***			
Gene	*p-adj*	*FC*	*Mean expression*[*]	*p-adj*	*FC*	*Mean expression*[*]	*Chrom*	*Start*	*End*	*Human name*
Ppp1r1b	2.1E-66	-20.61	1,333	2.6E-96	-26.88	2,097	11	98,348,404	98,357,796	PPP1R1B
Cml3	3.9E-03	-6.18	432	1.6E-68	-11.79	525	6	85,732,513	85,765,754	NAT8
Sirt3	3.5E-67	-6.00	1,519	1.4E-77	-5.46	1,995	7	140,863,670	140,882,309	SIRT3
Fah	1.6E-21	-5.79	5,895	3.6E-111	-4.32	7,337	7	84,585,159	84,606,722	FAH
Gcdh	1.3E-40	-5.08	7,306	1.3E-139	-4.58	10,056	8	84,886,393	84,893,921	GCDH
Hnf4a	1.7E-08	-5.01	15,950	9.1E-40	-4.13	21,250	2	163,506,808	163,572,910	HNF4A
Klf15	1.7E-56	-4.42	1,805	6.3E-10	-2.98	1,975	6	90,462,576	90,475,238	KLF15
Lhx1	1.2E-23	-4.24	635	7.7E-38	-3.67	917	11	84,518,284	84,525,535	LHX1
Nr0b2	1.9E-05	-4.01	31	8.7E-09	-5.51	49	4	133,553,376	133,556,536	NR0B2
Ar	4.4E-34	-3.76	2,420	2.7E-24	-2.80	2,500	X	98,149,721	98,323,215	AR
Hmx2	2.6E-14	-3.74	177	3.9E-09	-2.92	297	7	131,548,773	131,558,014	HMX2
Ppara	3.7E-15	-3.73	1,635	1.8E-13	-3.34	2,574	15	85,734,983	85,802,819	PPARA
Maf	2.7E-26	-3.37	6,289	1.8E-17	-2.82	8,918	8	115,682,942	115,707,794	MAF
Esr1	5.3E-22	-3.24	311	1.8E-10	-2.83	401	10	4,611,593	5,005,614	ESR1
Vdr	7.5E-25	-3.16	4,476	2.3E-36	-3.33	6,429	15	97,854,425	97,910,630	VDR
Ppargc1a	1.9E-22	-3.13	2,701	3.1E-38	-3.34	3,896	5	51,454,250	51,567,726	PPARGC1A
Jun	3.7E-17	2.66	2,691	3.2E-07	2.64	2,573	4	95,049,034	95,052,222	JUN
Pax8	8.6E-17	2.70	5,166	6.4E-44	3.21	4,895	2	24,420,560	24,475,599	PAX8
Trp53	1.2E-48	3.20	1,840	2.1E-12	2.80	1,586	11	69,580,359	69,591,873	TP53
Sqstm1	8.3E-18	3.25	29,263	4.9E-08	2.74	24,717	11	50,199,366	50,210,827	SQSTM1
Stat3	4.0E-109	3.37	6,849	7.4E-09	3.09	6,825	11	100,885,098	100,939,540	STAT3
Relb	6.5E-54	3.93	492	1.4E-06	2.75	462	7	19,606,217	19,629,438	RELB
Cebpb	8.5E-30	5.54	289	2.5E-18	5.25	214	2	167,688,915	167,690,418	CEBPB
Fos	8.9E-17	5.74	190	1.8E-24	6.29	157	12	85,473,890	85,477,273	FOS
Fosl2	3.9E-78	6.83	3,038	1.7E-63	6.00	2,929	5	32,135,801	32,157,842	FOSL2
Fosb	8.3E-27	7.60	44	3.4E-10	6.46	47	7	19,302,696	19,310,051	FOSB
Hesx1	1.6E-79	8.44	121	1.5E-39	10.54	87	14	27,000,362	27,002,329	HESX1
Egr1	9.5E-46	9.22	1,386	4.0E-13	8.14	1,373	18	34,859,823	34,864,984	EGR1
Atf3	3.2E-32	11.77	610	6.6E-44	16.21	414	1	191,170,296	191,218,039	ATF3
Bcl3	4.8E-83	14.17	651	1.7E-06	8.34	490	7	19,808,462	19,822,770	BCL3
Myc	4.0E-60	17.18	2,146	1.5E-153	15.91	1,506	15	61,985,341	61,990,374	MYC
Sox9	3.1E-69	20.70	403	8.8E-84	16.47	279	11	112,782,224	112,787,760	SOX9
Egr2	1.6E-50	30.93	200	4.8E-66	25.20	183	10	67,535,475	67,542,188	EGR2
Hmga2	5.2E-28	35.03	201	4.5E-50	31.15	135	10	120,361,275	120,476,469	HMGA2

[*]Mean calculated as average expression from samples included in each the comparison.

RNA-seq identified transcription factors driving the gene expression response to AKI. We detected, a total of 48 up-regulated and 41 down-regulated mRNAs coding for transcription factors, from which 33 were found to be disease term associated (Table 3; Figure 4).

3.4.3. Mechanisms of AKI activation driven by expression of Gdf15

Genetic response specific to Gdf15 expression was modulated by 100 differentially expressed genes (p-adj < 0.1, FC > 1.2) (Figure 2). Overall, 45 genes were up-regulated, and 55 were down-regulated in GDF15-KO-AKI with inversely related pattern of expression when compared to the WT-AKI/WT-Control comparison.

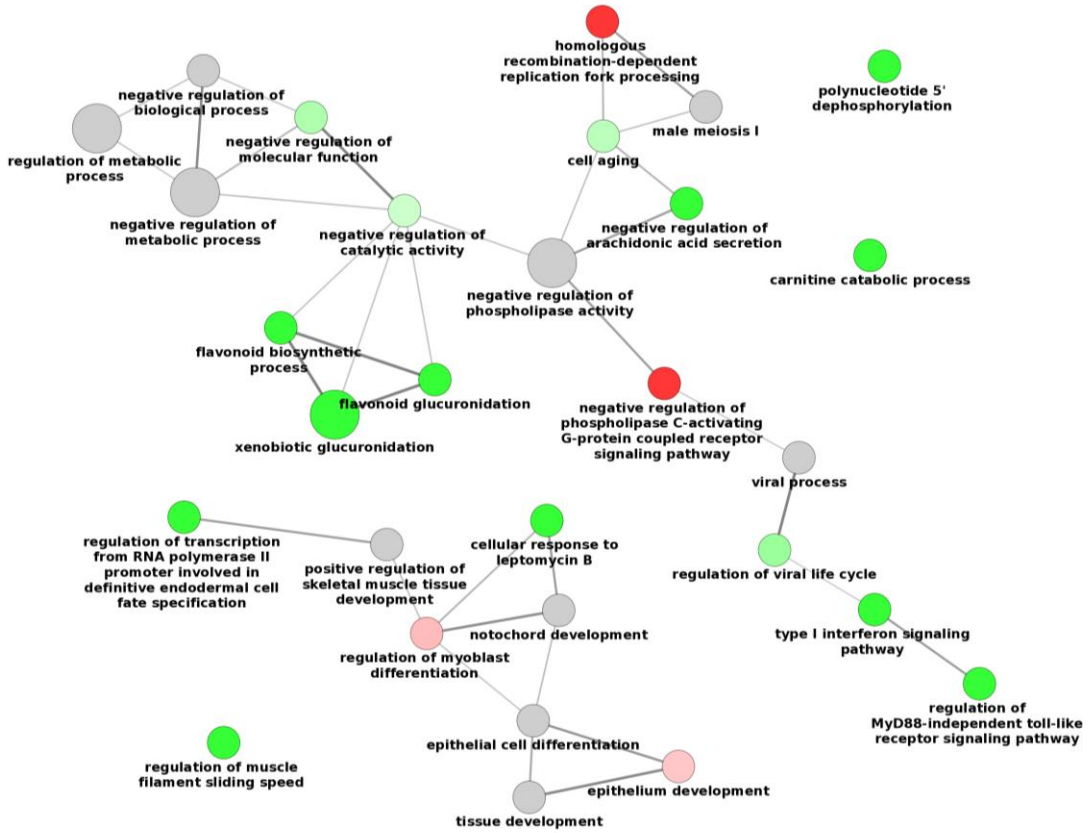


Figure 6: Enriched Gene Ontology Biological Processes lead by differentially expressed genes that are associated with a Gdf15 driven AKI response. Circles represent enriched terms that have an up-regulated (red), down-regulated (green), or no clear up- or down-regulated tendency (grey) in GDF15-KO-AKI. Terms have been clustered by term or process similarities and only the description of the major processes that define the clusters are shown (text). GO parent-child relationships between terms are shown as lines.

Biological processes driven by Gdf15 associated gene expression showed that up-regulated genes in GDF15-KO-AKI mice were related to processes such as epithelium development, tissue and cell morphogenesis, elastin biosynthetic process, positive regulation of microtubule motor activity, and regulation of membrane depolarization among other biological processes

(Figure 5; Supp. Figure S10). Down-regulated genes showed to be related to biological processes such as regulation of transcription from RNA polymerase II promoter involved in definitive endodermal cell fate specialization, xenobiotic gluconoridation, negative regulation of catalytic activity and arachidonic acid secretion, response to interferon and type I interferon signaling, endocytosis, and cell aging among other immune system, catalytic and metabolic biological processes (Figure 5; Supp. Figure S11).

Assessment of disease associated genes linked to Gdf15 expression revealed a total of 17 genes that were found differentially expressed with inversed regulation GDF15-KO and WT comparisons (Figure 6). No up-regulated and disease associated gene was found in an enriched and up-regulated biological processes (Figure 6). However, some down-regulated and disease associated genes were fount associated with down-regulated processes such as negative regulation of molecular function and catalytic activity, type I interferon signaling, and cell aging.

Table 4: Transcription factors driving the *Gdf15*-associated AKI response.

	GDF15-KO changes in AKI			***WT change in AKI***			***Genomic information***			
Gene*	*p-adj*	*FC*	*Mean expression*	*p-adj*	*FC*	*Mean expression*	*Chrom*	*Start*	*End*	*Human name*
Irf7	7.7E-03	-4.69	1,679	1	1.14	733	chr7	141,262,706	141,266,481	IRF7
Id3	3.6E-10	-2.37	2,021	1	1.15	1,861	chr4	136,143,497	136,145,755	ID3
Id1	1.2E-03	-1.98	1,095	1	1.60	913	chr2	152,736,251	152,737,410	ID1
Peg3	1.0E-07	-1.77	1,115	1	1.12	1,763	chr7	6,703,892	6,730,431	PEG3
Ebf1	1.8E-10	-1.60	368	1	1.00	466	chr11	44,617,317	45,008,091	EBF1
Sox17	6.2E-03	-1.30	363	1	1.13	448	chr1	4,490,931	4,497,354	SOX17
Zfp729b	1	1.02	511	6.1E-03	-1.51	686	chr13	67,589,443	67,609,707	ZNF729
4930522-L14Rik	1	1.10	107	7.0E-02	-1.86	154	chr5	109,735,990	109,751,886	ZNF788
Zfp26	6.7E-04	1.25	772	1	-1.00	966	chr9	20,432,972	20,460,162	
Irf6	2.4E-04	1.42	1,077	1	-1.02	1,340	chr1	193,153,111	193,172,023	IRF6
Zfp867	3.1E-02	1.42	210	1	-1.03	259	chr11	59,461,197	59,472,474	
Bcl6b	1.2E-03	1.57	341	1	-1.10	394	chr11	70,224,128	70,229,798	BCL6B
Rsl1	4.0E-11	1.67	175	1	-1.07	186	chr13	67,173,207	67,183,126	

**Bold indicates a previous association with renal disease terms.*

Evaluation of genetic drivers of AKI response identified 13 transcription factors associated to Gdf15 expression, from which, 8 were previously associated with AKI disease-terms (Table 4). Up-regulated transcription factors were found enriching biological processes such as tissue development (Figure 6). Whereas down-regulated transcription factors were enriching biological processes such as type I interferon signalling and negative regulation of molecular function (Figure 6).

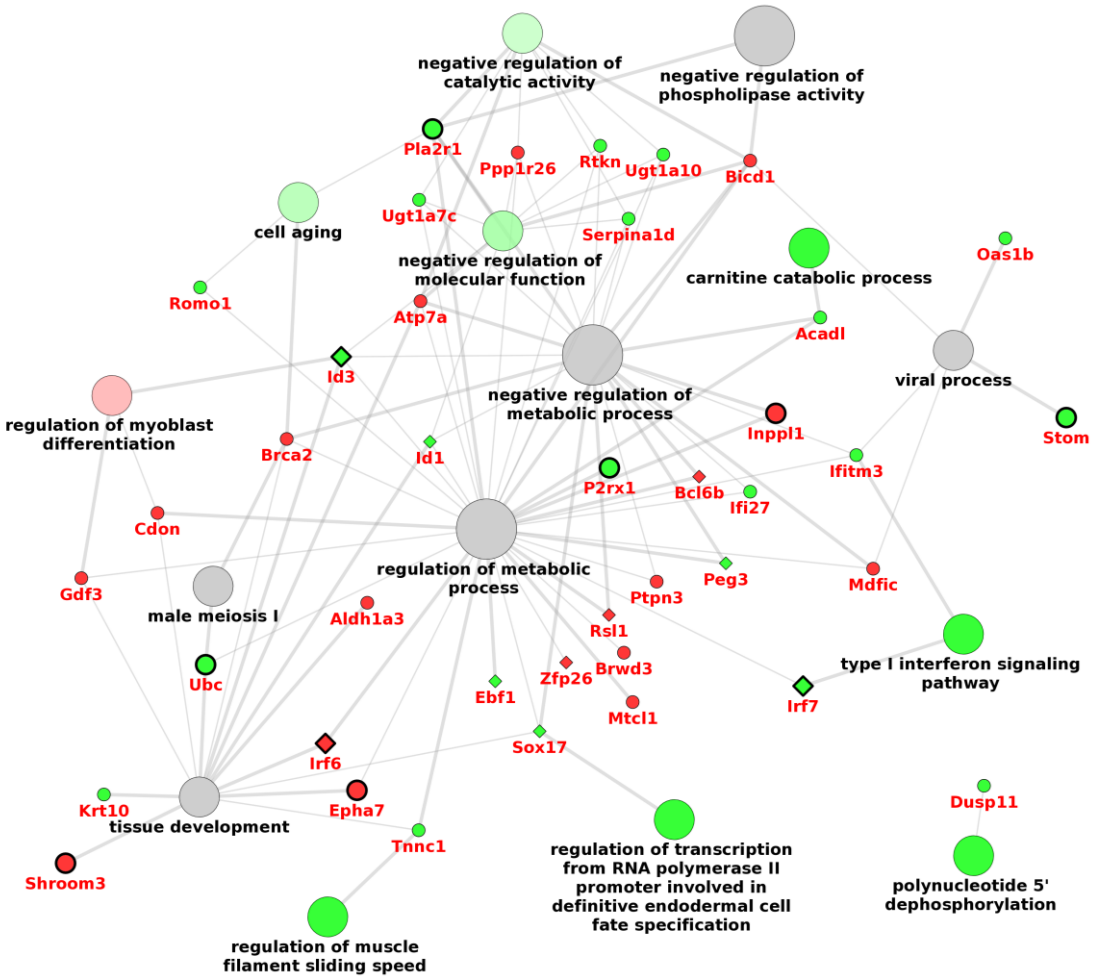


Figure 7: Enriched Gene Ontology Biological Processes lead by DE and disease associate genes with transcription factor activity that are associated to Gdf15 expression. Circles represent enriched terms that shown to be up-regulated (red), down-regulated (green), or no clear up- or down-regulated tendency (grey), using GDF15-KO-AKI mice as reference for up- or down- regulation. Terms have been clustered by term or process similarities and only the description of the major process that defines the cluster is shown (text). GO parent-child relationships between terms are shown as lines. Smaller circles represent genes enriching the annotated processes including disease associated genes (black border), and transcription factors genes (diamonds).

3.4.4. Biological pathways modulated by a transcription factor-driven AKI response

In the context of curated biological pathways, the assessment of a generic and a *Gdf15*-driven AKI associated genes enriched 71 biological pathways, 35 being up-regulated, and 35 down-regulated (Table 5). In addition, 24 up-regulated pathways included identified transcription factors, 4 pathways with both AKI and *Gdf15* associated transcription factors, 19 with AKI-only associated transcription factors, and 1 with *Gdf15*-only associated transcription factors. The down-regulated pathways included 7 pathways with AKI-only associated transcription factors. Only one pathway was enriched using both up- and down- regulated genes and showed no clear trend of being driven by up- or down-regulated expression.

Table 5: Enriched pathways in AKI.

	Pathway names	*Genes* (r/n)*	*%*	*Z (all)*	*p-value* (all)*	*Z (up)*	*p-value* (up)*	*Z (down)*	*p-value* (down)*	*TF* (Gdf15/AKI)*
Up-regulated Pathways	Cytoplasmic Ribosomal Proteins	76/78	97.4	7.61	**0**	12.5	**0**	-4.7	0	0/0
	TNF-alpha NF-kB Signaling Pathway	129/179	72.1	4.7	**0**	8.09	**0**	-3.31	0.002	3/23.1
	mRNA processing	255/436	58.5	1.57	0.114	7.91	**0**	-6.77	0	0/15.4
	Proteasome Degradation	43/58	74.1	2.96	**0.001**	6.54	**0**	-3.65	0	0/0
	DNA Replication	33/41	80.5	3.3	**0.001**	6.51	**0**	-3.22	0.001	0/46.2
	Translation Factors	34/48	70.8	2.23	**0.022**	5.34	**0**	-3.2	0.001	0/0
	G1 to S cell cycle control	45/61	73.8	2.98	**0.005**	5	**0**	-1.96	0.064	0/38.5
	EGFR1 Signaling Pathway	115/174	66.1	3.02	**0.006**	4.46	**0**	-1.33	0.192	0/57.7
	Podocyte prot-prot interactions (PodNet)	200/313	63.9	3.3	**0**	4.09	**0**	-0.59	0.554	0/3.8
	Alpha6-Beta4 Integrin Signaling Pathway	46/67	68.7	2.28	**0.016**	4.01	**0**	-1.71	0.085	0/7.7
	Podocyte prot-prot interactions, (XPodNet) **	472/813	58.1	1.97	0.059	3.6	**0**	-1.62	0.111	0/23.1
	miRNA regulation of DNA Damage Response	41/66	62.1	1.18	0.222	3.59	**0**	-2.53	0.01	0/30.8
	Nucleotide Metabolism	16/19	84.2	2.57	**0.006**	3.51	**0.001**	-0.82	0.432	0/0
	p38 MAPK Signaling Pathway	24/34	70.6	1.84	**0.068**	3.49	**0.001**	-1.65	0.089	0/15.4
	Focal Adhesion	108/185	58.4	0.97	0.332	3.45	**0**	-2.63	0.01	0/7.7
	TGF Beta Signaling Pathway	32/51	62.8	1.13	0.239	3.37	**0.001**	-2.35	0.013	0/30.8
	Integrin-mediated Cell Adhesion	59/99	59.6	0.95	0.329	3.32	**0.001**	-2.51	0.01	0/0
	Primary FSGS	46/69	66.7	1.98	**0.044**	3.27	**0**	-1.25	0.228	1.5/0
	Hypertrophy Model	13/18	72.2	1.48	0.121	3.26	**0**	-1.82	0.066	0/7.7
	Regulation of Actin Cytoskeleton	81/145	55.9	0.23	0.818	3.06	**0.003**	-3.06	0.001	0/0
	IL-1 Signaling Pathway	24/37	64.9	1.22	0.236	3.02	**0.004**	-1.86	0.066	0/0
	Lung fibrosis	36/59	61.0	0.95	0.343	3	**0.002**	-2.16	0.032	0/3.8
	Homologous recombination	9/13	69.2	1.04	0.278	2.97	**0.004**	-2.01	0.03	0/0
	Toll Like Receptor signalling	21/33	63.6	1.01	0.307	2.91	**0.003**	-1.98	0.035	0/0
	IL-3 Signaling Pathway	58/97	59.8	0.98	0.343	2.83	**0.005**	-1.94	0.057	3/7.7
	PluriNetWork	140/273	51.3	-1.24	0.197	2.81	**0.007**	-4.51	0	1.5/84.6
	MAPK signaling pathway	85/158	53.8	-0.29	0.757	2.75	**0.014**	-3.33	0	0/34.6
	FAS-pathway/Stress-induction of HSP regulation	22/37	59.5	0.56	0.597	2.66	**0.004**	-2.24	0.021	0/7.7
	EBV LMP1 signaling	14/21	66.7	1.08	0.327	2.58	**0.006**	-1.53	0.131	0/0
	Insulin Signaling	95/156	60.9	1.53	0.14	2.53	**0.008**	-0.96	0.334	0/30.8
	IL-6 signaling Pathway	60/99	60.6	1.15	0.259	2.44	**0.014**	-1.31	0.204	0/100
	Mismatch repair	6/9	66.7	0.71	0.425	2.3	**0.013**	-1.68	0.075	0/0
	Apoptosis	44/82	53.7	-0.23	0.811	2.27	**0.031**	-2.74	0.006	6.1/23.1
	Signaling of Hepatocyte Growth Factor Receptor	19/34	55.9	0.11	0.889	2.01	**0.049**	-2.05	0.041	0/23.1
Down-regulated pathways	Electron Transport Chain	89/100	89.0	6.92	**0**	-6.79	0	15.48	**0**	0/0
	Oxidative phosphorylation	54/59	91.5	5.68	**0**	-5.2	0	12.3	**0**	0/0
	TCA Cycle	26/30	86.7	3.51	**0**	-3.3	0.001	7.69	**0**	0/0
	One carbon metabolism and related pathways	37/49	75.5	2.91	**0.006**	-3.18	0.001	6.87	**0**	0/0
	PPAR signaling pathway	52/74	70.3	2.67	**0.008**	-3.05	0.003	6.45	**0**	0/3.8
	Fatty Acid Beta Oxidation	25/33	75.8	2.41	**0.014**	-2.75	0.004	5.81	**0**	0/7.7
	Amino Acid metabolism	67/92	72.8	3.48	**0.002**	-1.51	0.136	5.72	**0**	0/3.8
	Mitochondrial LC-Fatty Acid Beta-Oxidation	14/16	87.5	2.62	**0.008**	-2.15	0.027	5.41	**0**	0/0
	Glutathione and one carbon metabolism	24/31	77.4	2.53	**0.011**	-2.2	0.028	5.35	**0**	0/0
	Fatty acid oxidation	9/10	90	2.23	**0.023**	-2.13	0.035	4.93	**0**	0/0
	Metapathway biotransformation	79/132	59.9	1.16	0.247	-3.26	0.001	4.9	**0**	0/0
	Tryptophan metabolism	27/42	64.3	1.23	0.217	-3.04	0.003	4.74	**0**	0/15.4
	Selenium Micronutrient Network	18/23	78.3	2.26	**0.027**	-1.88	0.067	4.68	**0**	0/0
	Glutathione metabolism	16/19	84.2	2.57	**0.013**	-1.45	0.144	4.58	**0**	0/0
	Fatty Acid Biosynthesis	18/22	81.8	2.54	**0.008**	-1.32	0.176	4.41	**0**	0/0
	Folic Acid Network	19/23	82.6	2.68	**0.011**	-0.98	0.351	4.19	**0**	0/0
	Synthesis and Degradation of Ketone Bodies	5/5	100	2.03	**0.043**	-1.51	0.122	4.01	**0.001**	0/0
	Glycolysis and Gluconeogenesis	29/46	63.0	1.11	0.287	-2.35	0.022	3.85	**0**	0/0
	One Carbon Metabolism	21/29	72.4	1.9	**0.049**	-1.22	0.234	3.55	**0.001**	0/0
	Glucocorticoid & Mineralcorticoid Metabolism	7/11	63.6	0.58	0.566	-2.23	0.036	3.11	**0.002**	0/0
	Arachidonate Epoxygenase Epoxide Hydrolase	3/3	100	1.57	0.113	-1.17	0.241	3.1	**0.001**	0/0
	Statin Pathway	13/17	76.5	1.79	0.086	-0.68	0.517	2.83	**0.006**	0/0
	Steroid Biosynthesis	8/12	66.7	0.82	0.437	-1.71	0.096	2.82	**0.001**	0/0
	Alanine and aspartate metabolism	8/12	66.7	0.82	0.423	-1.71	0.102	2.82	**0.003**	0/0
	Nuclear Receptors	19/38	50.0	-0.61	0.55	-3.11	0.001	2.67	**0.003**	0/53.8
	Eicosanoid Synthesis	11/18	61.1	0.53	0.588	-1.84	0.052	2.62	**0.005**	0/0
	Methylation	7/8	87.5	1.85	**0.048**	-0.38	0.687	2.58	**0.005**	0/0
	Glucuronidation	9/13	69.2	1.04	0.318	-1.23	0.234	2.55	**0.008**	0/0
	Amino acid conjugation of benzoic acid	2/2	100	1.28	0.224	-0.95	0.348	2.53	**0.008**	0/0
	Selenium metabolism/Selenoproteins	34/46	73.9	2.6	**0.012**	0.53	0.598	2.46	**0.012**	0/15.4
	Polyol pathway	3/4	75.0	0.81	0.396	-1.35	0.188	2.41	**0.013**	0/0
	Oxidative Stress	23/28	82.1	2.9	**0.003**	0.93	0.369	2.38	**0.016**	0/11.5
	Heme Biosynthesis	5/9	55.6	0.04	0.992	-2.02	0.052	2.24	**0.021**	0/0
	Aflatoxin B1 metabolism	3/5	60.0	0.23	0.867	-1.51	0.117	1.91	**0.042**	0/0
	Prostaglandin Synthesis and Regulation	21/29	72.4	1.9	**0.044**	1.19	0.222	0.92	0.317	0/0

* Bold: significant p-value < 0.05. Genes found differentially expressed in the pathway (r) or measured in the dataset (n). Percentage of differentially expressed transcription factors (TF) in the pathway. ** "...and expanded by STRING".

3.5. Discussion

Accumulated nitrogen metabolites (urea and creatinine as well as other unmeasured waste products), acids, potassium, and others are characteristics of the AKI syndrome among other effects such as decreased urine output (Bellomo et al., 2012). AKI commonly occurs in hospitalized or critically ill patients with a prevalence between 20% to 60% for patients at admission to the intensive care unit when sepsis is present (Bellomo et al., 2012). Toxin exposure, usually corresponding to nephrotoxic drugs, is among the three most common causes of AKI seen in humans (Yang et al., 2010). The other two major causes are ischemia, and urinary obstructions.

An accepted model for toxin exposure is folic acid nephropathy which has also been reported in humans (Metz-Kurschel et al., 1990). Acute folate nephropathy causes reversible increase in serum creatinine and urea, tubular cell death, compensatory tubular cell proliferation, activation of an inflammatory response and eventual progression to mild fibrosis (Fang et al., 2005; Doi et al., 2006; Ortega et al., 2006). Therefore, we employed a folic acid overdose induced animal model representative for toxic exposure AKI, in wild type and *Gdf15* deficient mice. In this study, we used an mRNA sequencing-based non-biased approach to elucidate driver gene expression determinants of biological processes and pathway mechanisms of AKI, as well as the *Gdf15*-associated tubular damage protection-related mechanisms.

Most of the pathways previously known to be commonly involved in AKI could be confirmed with our transcriptomics approach, including a locally activated coagulation system, leucocyte infiltrates, inflammation, injured endothelium and expressed adhesion molecules, release of cytokines, induced toll-like receptor, vasoconstriction, transport activity and homeostasis, fibrosis, necrosis, and apoptosis (Table 5) (Bellomo et al., 2012; Vallon, 2016). In addition, in most of these or related pathways we identified, possible transcription factors driving the gene expression response to AKI for both AKI- and *Gdf15*-associated responses (Table 5). Notably, transcription factors associated with *Gdf15* deficiency are driving up-regulated pathways such as *TNF-alpha NF-κB Signaling Pathway*, *Primary Focal Segmental Glomerulosclerosis (FSGS)*, *IL-3 Signaling Pathway*, and *Apoptosis* (Table 5), whereas no *Gdf15*-associated transcription factor was identified for down-regulated pathways (Table5), suggesting a primary role of *Gdf15* in the regulation of processes involved with proliferation, inflammation, necrosis, fibrosis, and apoptosis among other processes regulated

by these pathways. AKI also results in interstitial inflammation, development of fibrosis and the production of pro-fibrotic cytokines that depend on proximal tubule epithelial cells arrest in G2/M during the cell cycle (Yang et al., 2010). Bypassing the tubular cells G2/M arrest in murine AKI prevented the transition from AKI to CKD (Yang et al., 2010). Our results show an upregulation of cell cycle stage G2 genes (Supp. Figure S13). However, also other stages of the cell cycle showed upregulated genes (Supp. Figure S13). This is consistent with cell cycle arrest in some cells, thus promoting chronicity and cell cycle progression in other cells, resulting in possible recovery from injury. In addition, we observed that only AKI-associated transcription factors are modulating the up-regulated cell cycle pathway *G1 to S cell cycle control* (Table 5). However, in *Gdf15* deficient mice, with a more severe AKI, this upregulation was not observed, suggesting a role of *Gdf15* on the cell cycle arrest, directly or indirectly, that could have an impact in the protection or recovery after tubular damage (Yang et al., 2010).

Using a proteomics approach on the same AKI mouse model, Husi et al., (2013) reported that signal transduction cascades in AKI end in kidney apoptosis and necrosis through the *N*-methyl-D-aspartate (NMDA) receptor (Husi et al., 2013). mRNA-seq data correlated with many of the key upregulated proteins reported by Husi et al., 2013 (Supp. Figure S12). However, the expression of some genes was not similarly upregulated, or even inversely regulated compared to the proteins they encoded (Supp. Figure S12). Similar observations were previously reported by Poveda *et al.*, 2016, where NFκBiz, one of the most upregulated NFκB-related genes during AKI, showed decreased protein levels, thus favoring chemokine production (Poveda et al., 2016). Messenger-RNA/protein level inconsistencies can be explained by per-transcript differences in transcript/protein relative expression abundance of 17/50,000 on average, as well as different protein and transcript half-life values of average 46h and 9h, respectively (Schwanhäusser et al., 2011). Possibly, taking into consideration each individual transcript and protein half-life as a metabolic response factor, the differences observed between transcript and protein at the same sampling point may simply reflect the differences in their metabolic rates. Further work is still required to characterize the cellular status and changes from the induction of AKI and its damaging mechanisms to the kidney, until the recovery of the renal function at both the transcriptomic and proteomic level.

Overall, mRNA sequencing revealed the expression profiles associated with the generic AKI response of folic acid induced AKI mice, as well as the specific response of AKI as modified by *Gdf15* deficiency. Furthermore, we identified the transcription factors associated to each

of these transcriptomes that are driving the generic and *Gdf15*-specific AKI trnscriptomic response. This information may be used to design novel therapeutic approaches for AKI.

3.6. References

Adela R, Banerjee SK. 2015. GDF-15 as a Target and Biomarker for Diabetes and Cardiovascular Diseases: A Translational Prospective. J Diabetes Res 2015:490842.

Baek SJ, Okazaki R, Lee S-H, Martinez J, Kim J-S, Yamaguchi K, Mishina Y, Martin DW, Shoieb A, McEntee MF, Eling TE. 2006. Nonsteroidal anti-inflammatory drug-activated gene-1 over expression in transgenic mice suppresses intestinal neoplasia. Gastroenterology 131:1553–1560.

Bellomo R, Kellum JA, Ronco C. 2012. Acute kidney injury. Lancet Lond Engl 380:756–766.

Breit SN, Carrero JJ, Tsai VW-W, Yagoutifam N, Luo W, Kuffner T, Bauskin AR, Wu L, Jiang L, Barany P, Heimburger O, Murikami M-A, et al. 2012. Macrophage inhibitory cytokine-1 (MIC-1/GDF15) and mortality in end-stage renal disease. Nephrol Dial Transplant Off Publ Eur Dial Transpl Assoc - Eur Ren Assoc 27:70–75.

Doi K, Okamoto K, Negishi K, Suzuki Y, Nakao A, Fujita T, Toda A, Yokomizo T, Kita Y, Kihara Y, Ishii S, Shimizu T, et al. 2006. Attenuation of Folic Acid-Induced Renal Inflammatory Injury in Platelet-Activating Factor Receptor-Deficient Mice. Am J Pathol 168:1413–1424.

Fang T-C, Alison MR, Cook HT, Jeffery R, Wright NA, Poulsom R. 2005. Proliferation of bone marrow-derived cells contributes to regeneration after folic acid-induced acute tubular injury. J Am Soc Nephrol JASN 16:1723–1732.

Guide for the Care and Use of Laboratory Animals: Eighth Edition. 2011. Washington, D.C.: National Academies Press.

Hellemons ME, Mazagova M, Gansevoort RT, Henning RH, Zeeuw D de, Bakker SJL, Lambers-Heerspink HJ, Deelman LE. 2012. Growth-Differentiation Factor 15 Predicts Worsening of Albuminuria in Patients With Type 2 Diabetes. Diabetes Care 35:2340–2346.

Hsiao EC, Koniaris LG, Zimmers-Koniaris T, Sebald SM, Huynh TV, Lee SJ. 2000. Characterization of growth-differentiation factor 15, a transforming growth factor beta superfamily member induced following liver injury. Mol Cell Biol 20:3742–3751.

Husi H, Sanchez-Niño MD, Delles C, Mullen W, Vlahou A, Ortiz A, Mischak H. 2013. A combinatorial approach of Proteomics and Systems Biology in unravelling the mechanisms of acute kidney injury (AKI): involvement of NMDA receptor GRIN1 in murine AKI. BMC Syst Biol 7:110.

Kutmon M, Iersel MP van, Bohler A, Kelder T, Nunes N, Pico AR, Evelo CT. 2015. PathVisio 3: An Extendable Pathway Analysis Toolbox. PLOS Comput Biol 11:e1004085.

Kutmon M, Riutta A, Nunes N, Hanspers K, Willighagen EL, Bohler A, Mélius J, Waagmeester A, Sinha SR, Miller R, Coort SL, Cirillo E, et al. 2016. WikiPathways: capturing the full diversity of pathway knowledge. Nucleic Acids Res 44:D488–D494.

Li C, Wang J, Kong J, Tang J, Wu Y, Xu E, Zhang H, Lai M. 2016. GDF15 promotes EMT and metastasis in colorectal cancer. Oncotarget 7:860–872.

Mazagova M, Buikema H, Buiten A van, Duin M, Goris M, Sandovici M, Henning RH, Deelman LE. 2013. Genetic deletion of growth differentiation factor 15 augments renal damage in both type 1 and type 2 models of diabetes. Am J Physiol Renal Physiol 305:F1249-1264.

Metz-Kurschel U, Kurschel E, Wagner K, Aulbert E, Graben N, Philipp T. 1990. Folate nephropathy occurring during cytotoxic chemotherapy with high-dose folinic acid and 5-fluorouracil. Ren Fail 12:93–97.

Ortega A, Rámila D, Ardura JA, Esteban V, Ruiz-Ortega M, Barat A, Gazapo R, Bosch RJ, Esbrit P. 2006. Role of parathyroid hormone-related protein in tubulointerstitial apoptosis and fibrosis after folic acid-induced nephrotoxicity. J Am Soc Nephrol JASN 17:1594–1603.

Ortiz A, Husi H, Gonzalez-Lafuente L, Valiño-Rivas L, Fresno M, Sanz AB, Mullen W, Albalat A, Mezzano S, Vlahou T, Mischak H, Sanchez-Niño MD. 2017. Mitogen-Activated Protein Kinase 14 Promotes AKI. J Am Soc Nephrol JASN 28:823–836.

Poveda J, Sanz AB, Rayego-Mateos S, Ruiz-Ortega M, Carrasco S, Ortiz A, Sanchez-Niño MD. 2016. NFκBiz protein downregulation in acute kidney injury: Modulation of inflammation and survival in tubular cells. Biochim Biophys Acta 1862:635–646.

Sándor N, Schilling-Tóth B, Kis E, Benedek A, Lumniczky K, Sáfrány G, Hegyesi H. 2015. Growth Differentiation Factor-15 (GDF-15) is a potential marker of radiation response and radiation sensitivity. Mutat Res Genet Toxicol Environ Mutagen 793:142–149.

Schwanhäusser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W, Selbach M. 2011. Global quantification of mammalian gene expression control. Nature 473:337–342.

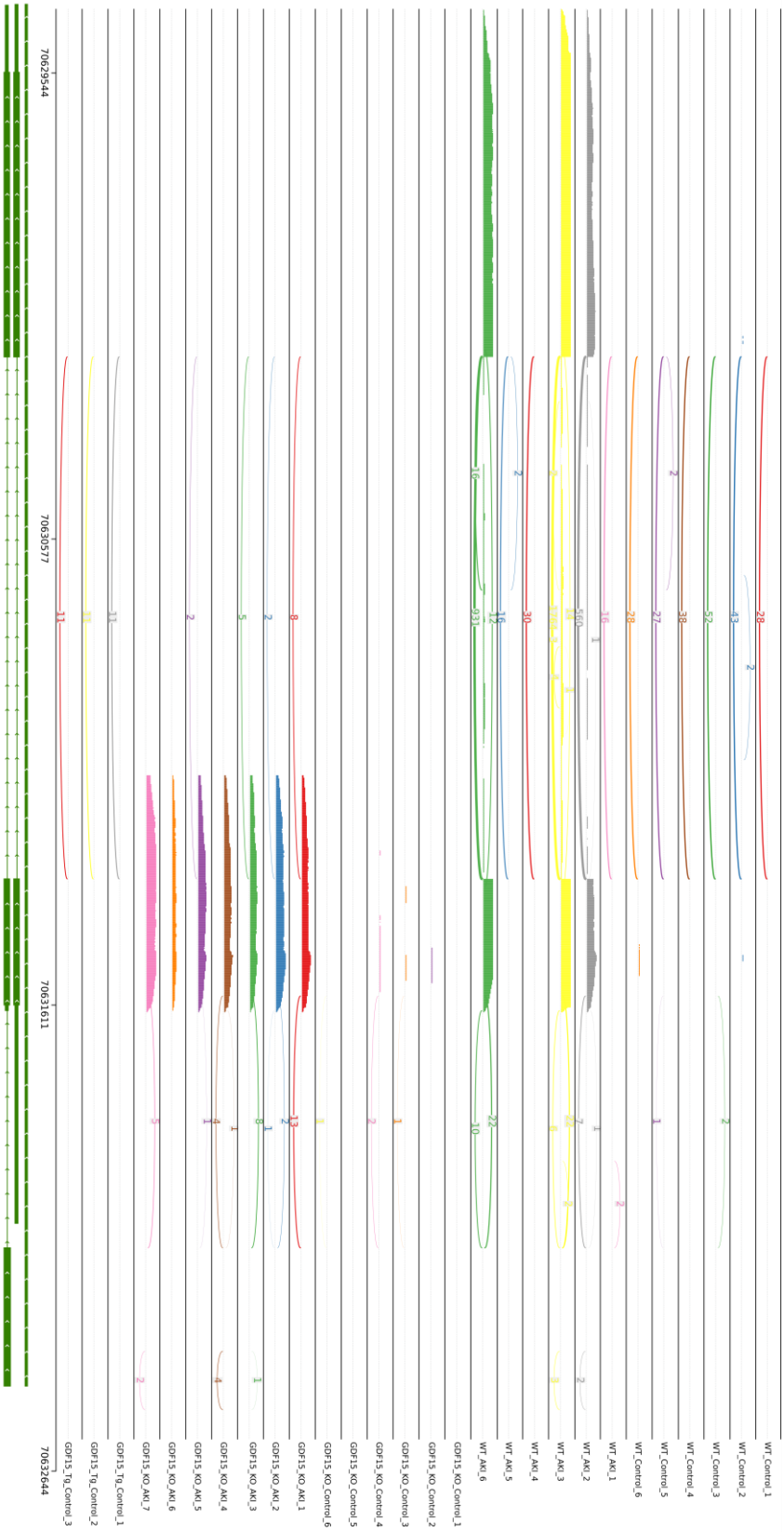
Vallon V. 2016. Tubular Transport in Acute Kidney Injury: Relevance for Diagnosis, Prognosis and Intervention. Nephron 134:160–166.

Yang L, Besschetnova TY, Brooks CR, Shah JV, Bonventre JV. 2010. Epithelial cell cycle arrest in G2/M mediates kidney fibrosis after injury. Nat Med 16:535–143.

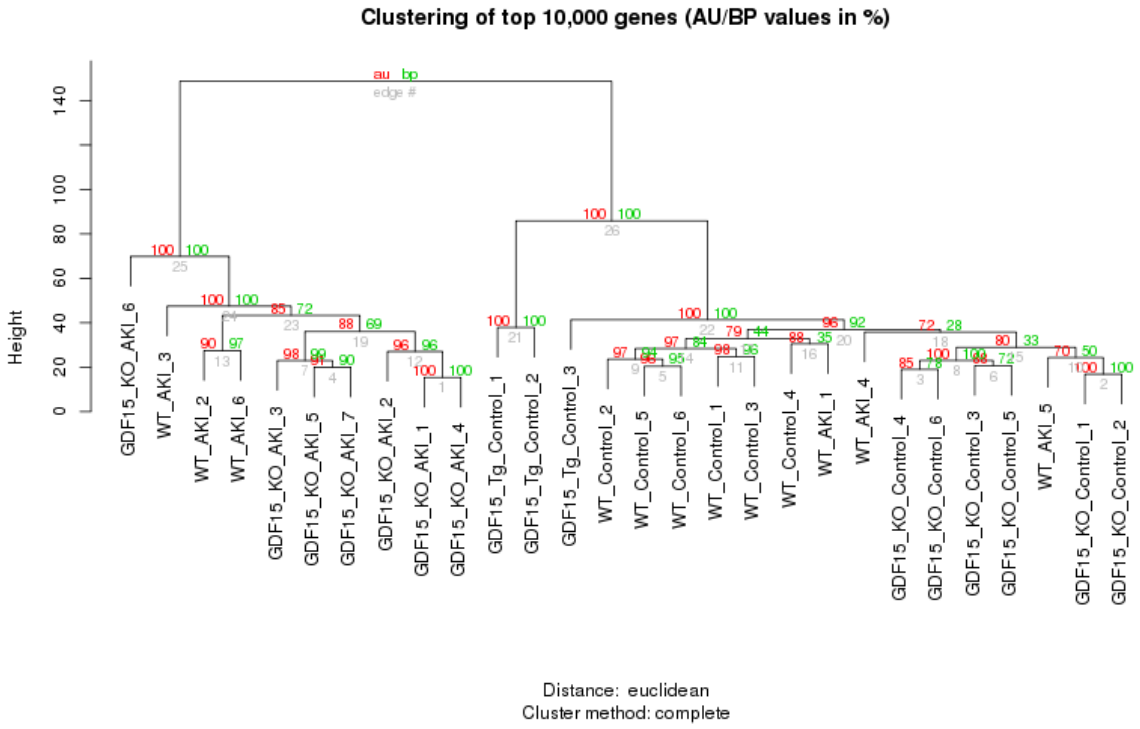
Yatsuga S, Fujita Y, Ishii A, Fukumoto Y, Arahata H, Kakuma T, Kojima T, Ito M, Tanaka M, Saiki R, Koga Y. 2015. Growth differentiation factor 15 as a useful biomarker for mitochondrial disorders. Ann Neurol 78:814–823.

Zhang H-M, Chen H, Liu W, Liu H, Gong J, Wang H, Guo A-Y. 2012. AnimalTFDB: a comprehensive animal transcription factor database. Nucleic Acids Res 40:D144–D149.

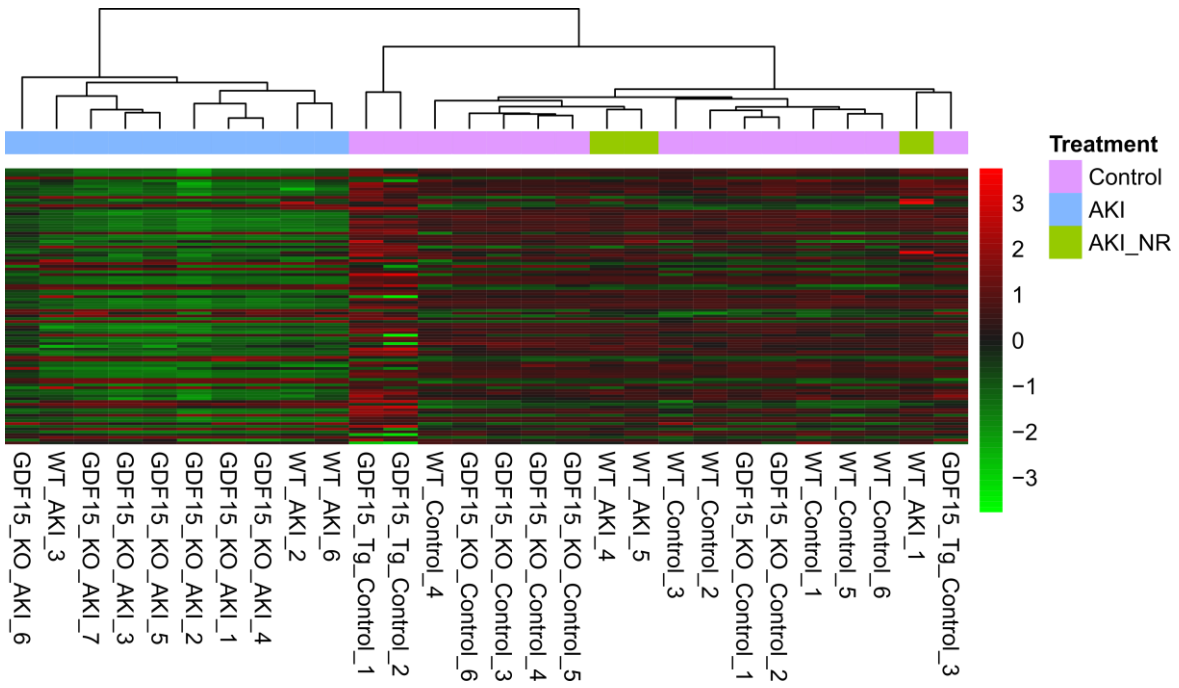
3.7. Annex II: Supplementary material of Chapter 3



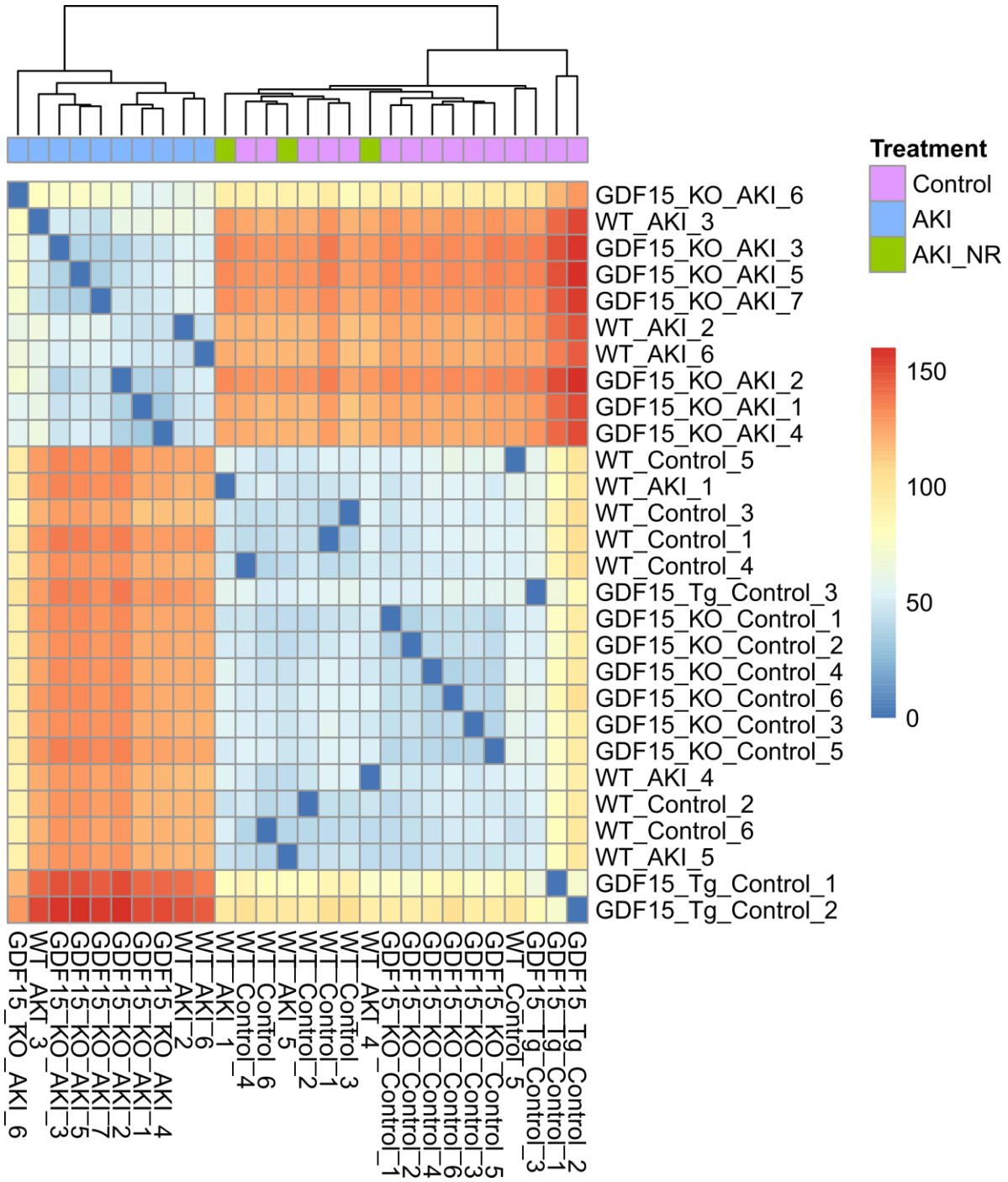
Supp. Figure S1: Sequence alignments of all RNA-seq samples, represented as coverage and spliced reads for the *Gdf15* gene region.



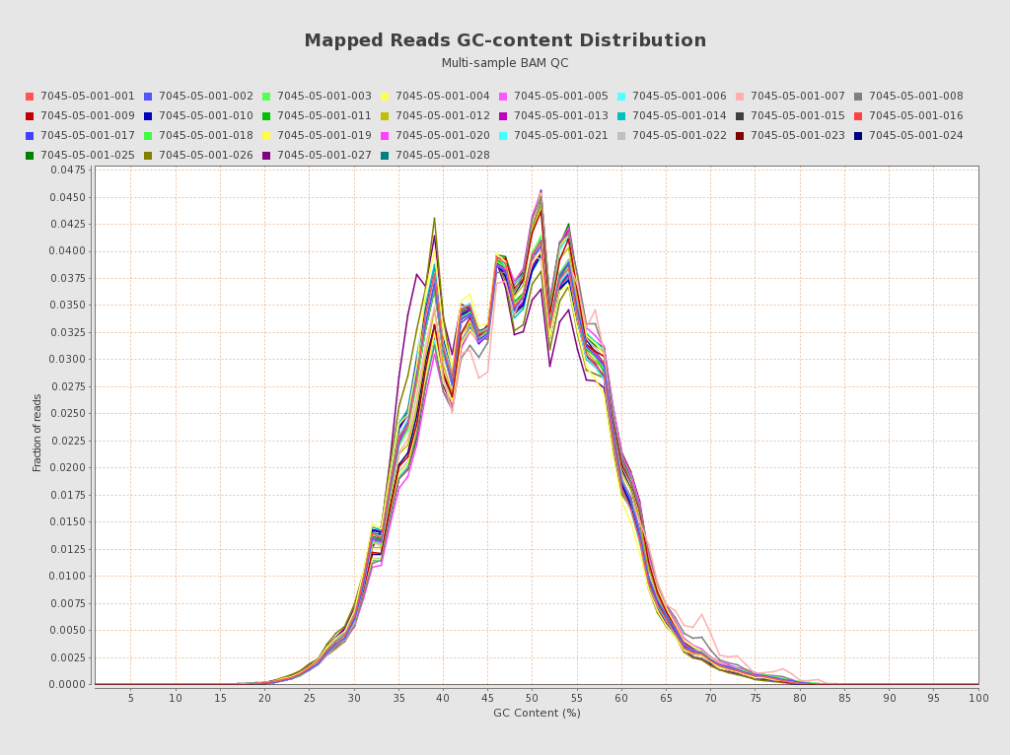
Supp. Figure S2: Unsupervised hierarchical clustering of samples based on the top 10,000 expressed genes. Cluster strengths shown as approximately unbiased (AU), and bootstrap probability (BP) p-values in red and green, respectively.



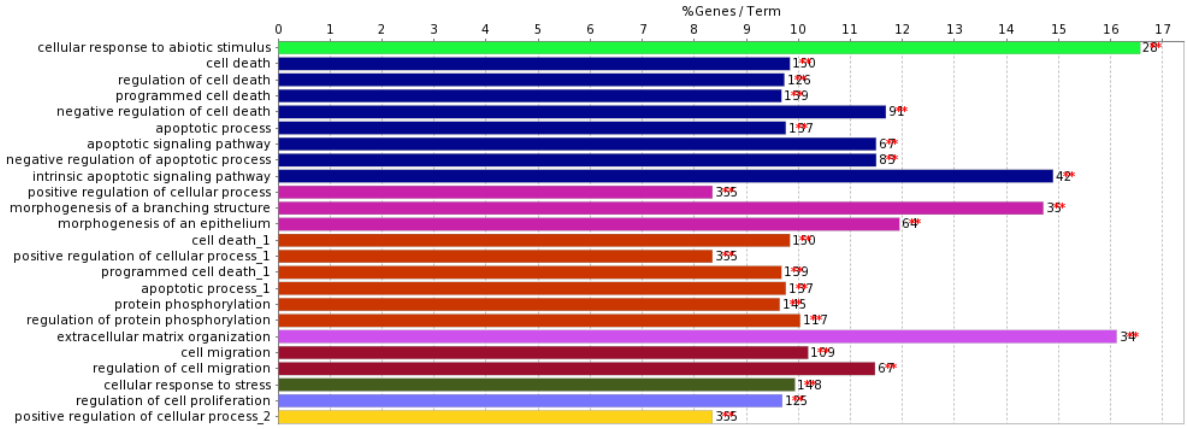
Supp. Figure S3: Heatmap of the top 100 expressed genes. Top clustered tree represents the grouping of each sample respective to the treatment received.



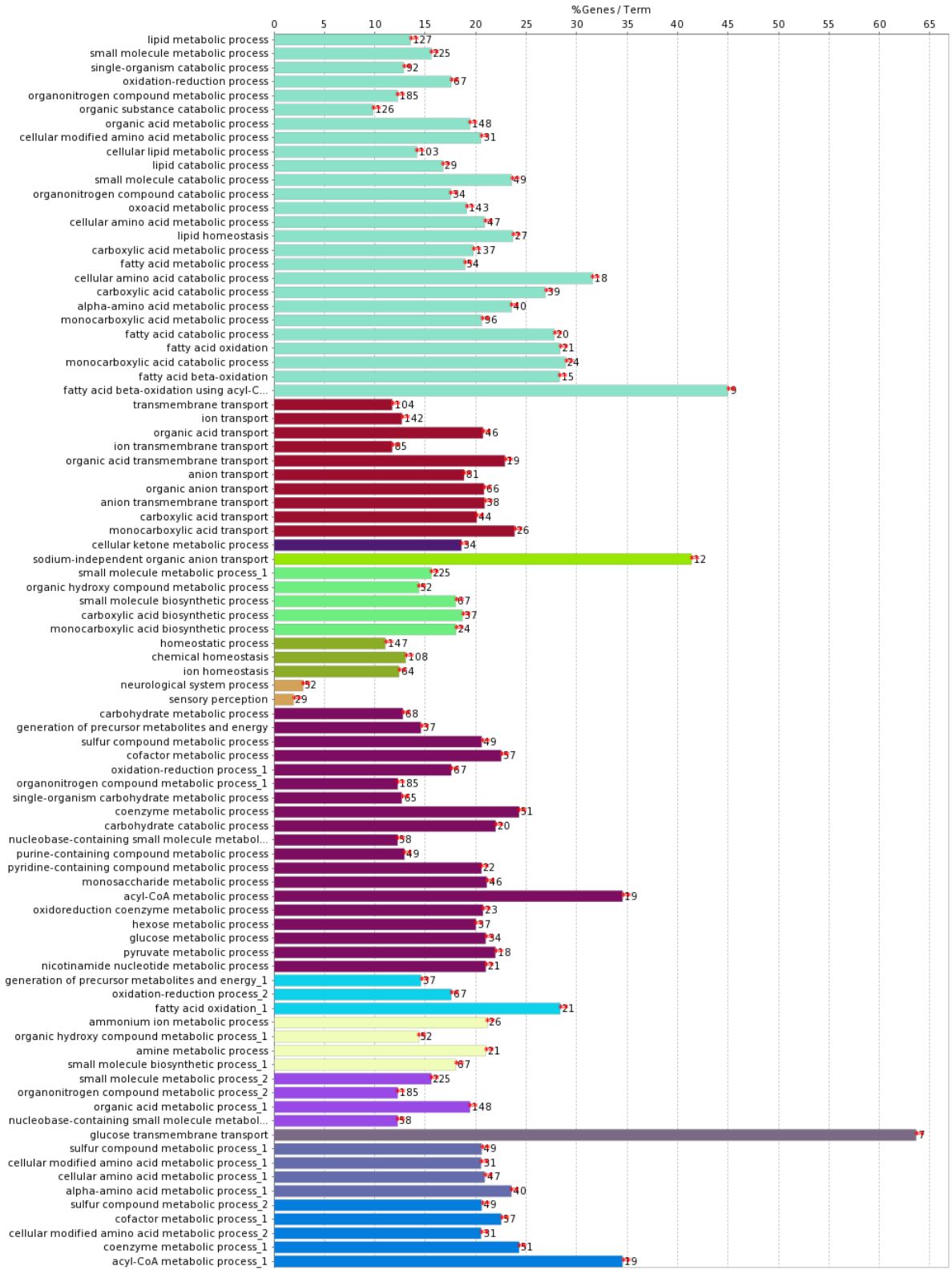
Supp. Figure S4: Heatmap of sample distances which can clearly discriminate between AKI responder and non-AKI responder samples.



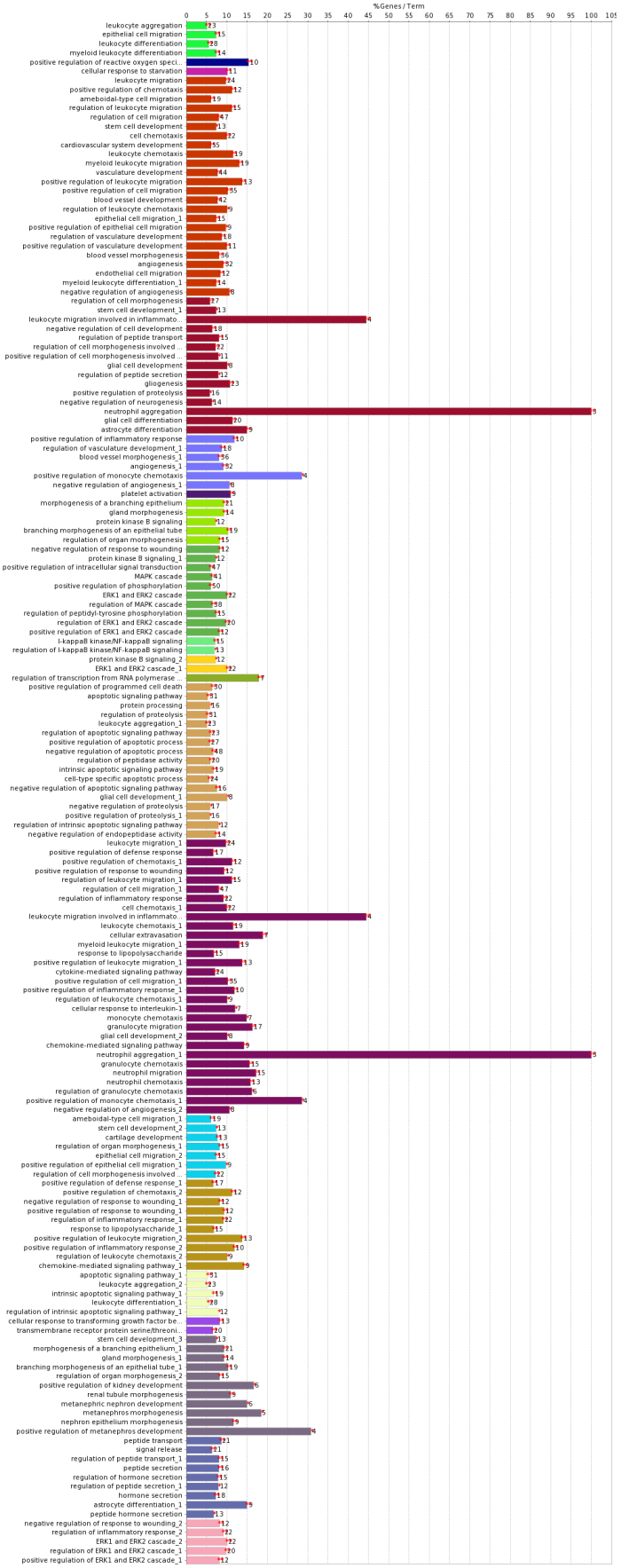
Supp. Figure S5: GC content distribution of all sequenced samples.



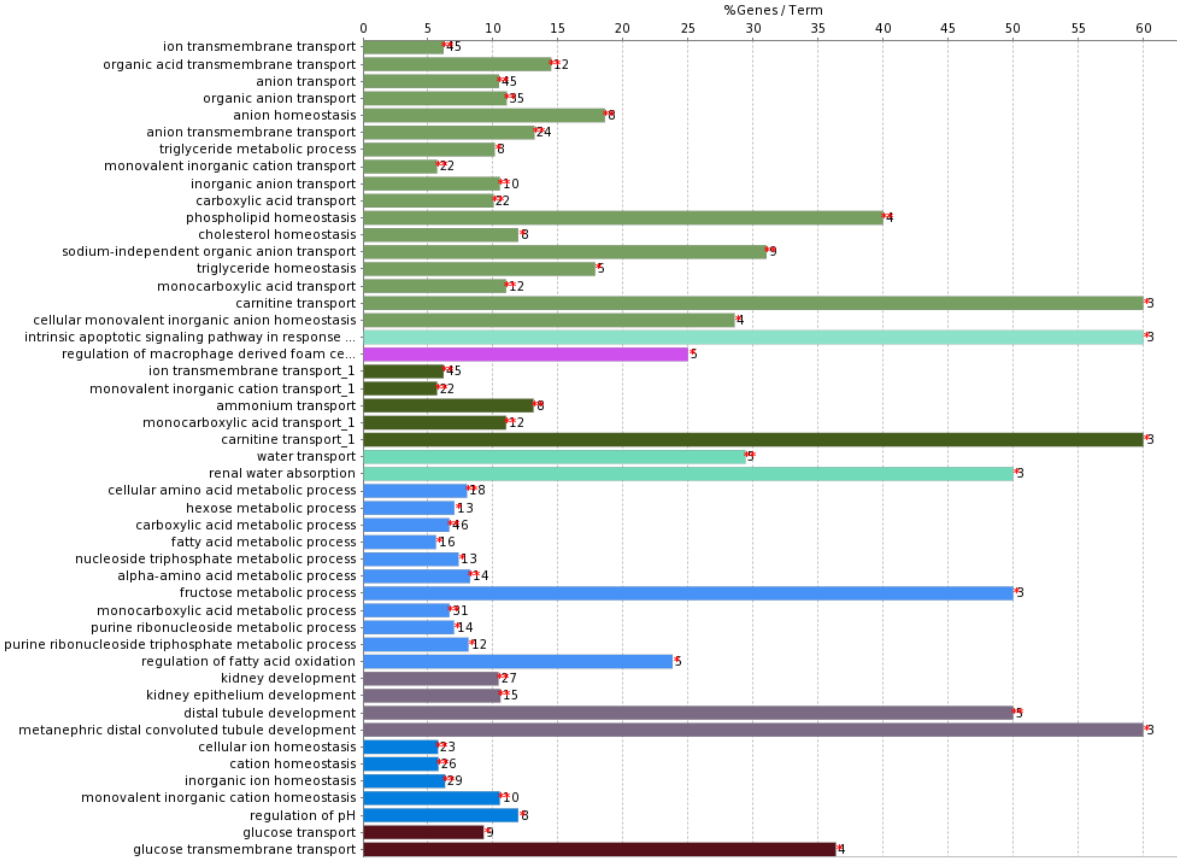
Supp. Figure S6: Gene count of a cluster of enriched processes that contain up-regulated genes in an AKI generic response.



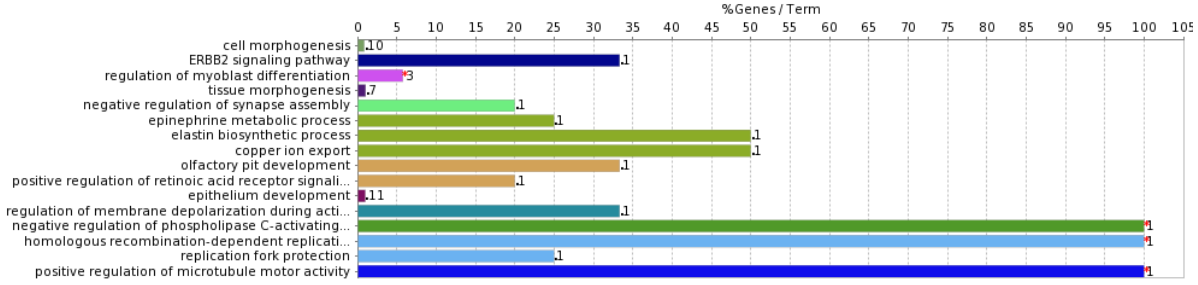
Supp. Figure S7: Gene count of a cluster of enriched processes that contain down-regulated genes in an AKI generic response.



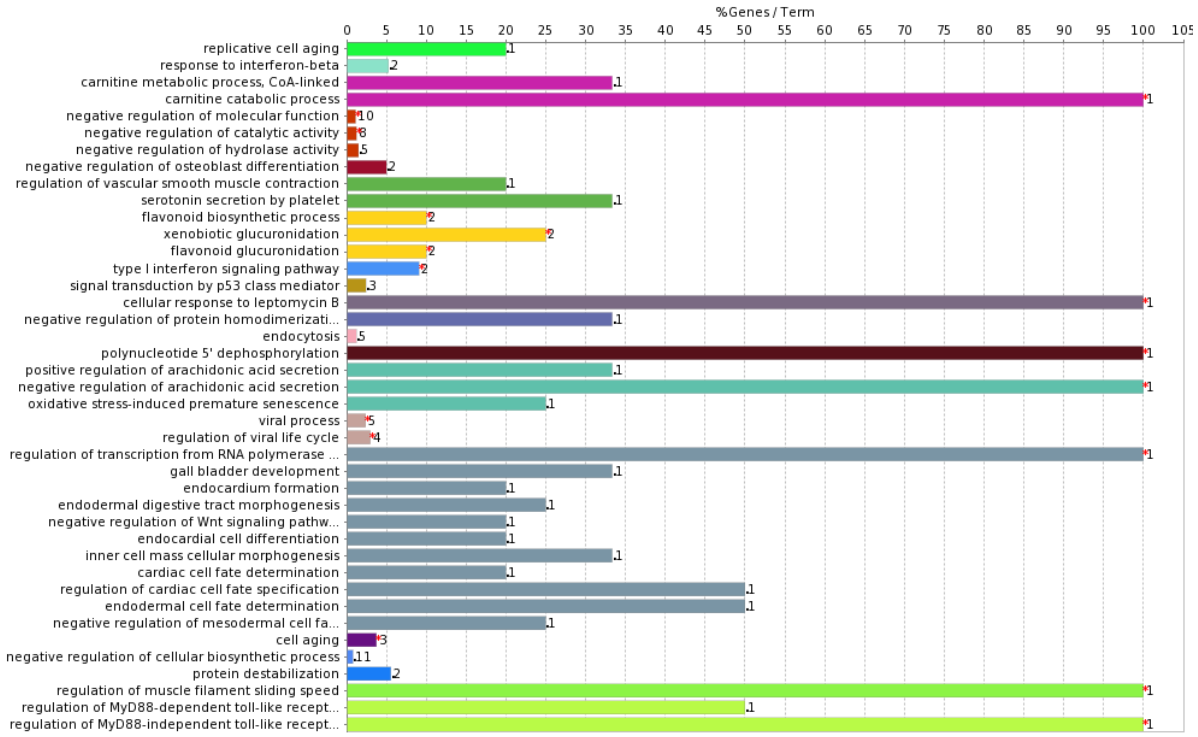
Supp. Figure S8: Gene count of a cluster of enriched processes that contain up-regulated genes in an AKI generic response that were previously associated with kidney disease terms.



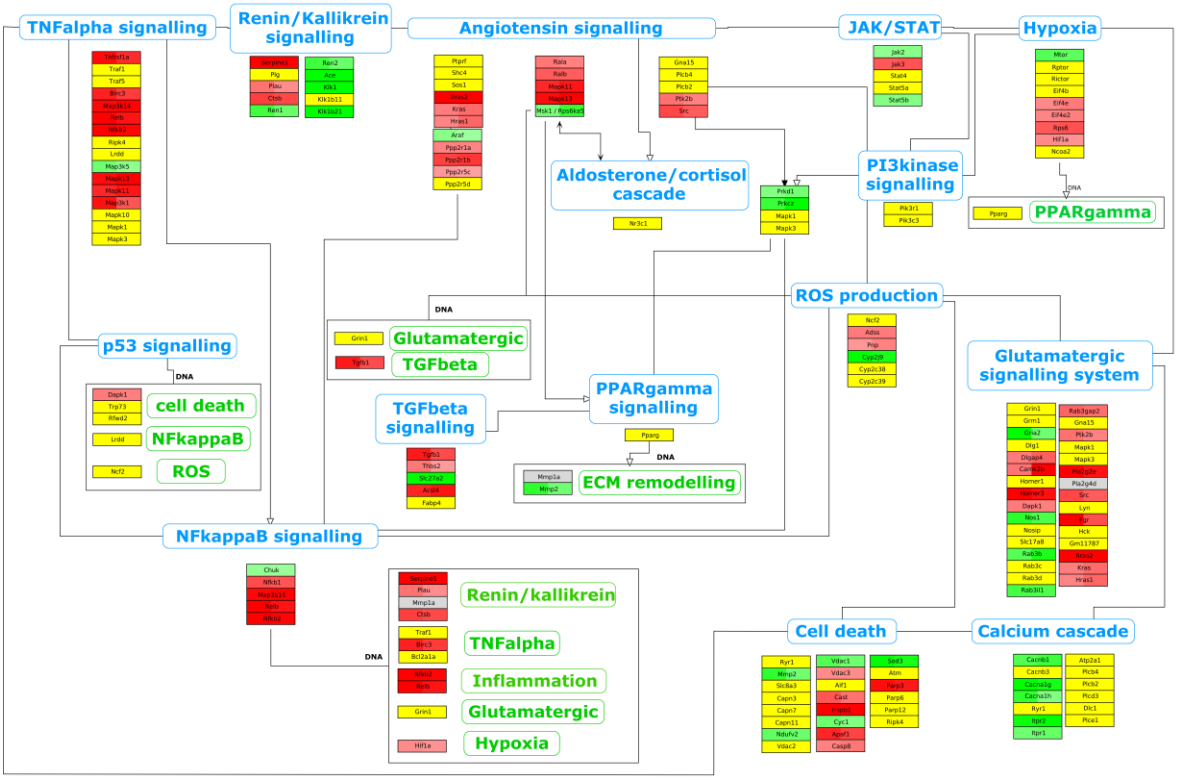
Supp. Figure S9: Gene count of a cluster of enriched processes that contain down-regulated genes in an AKI generic response that were previously associated with kidney disease terms.



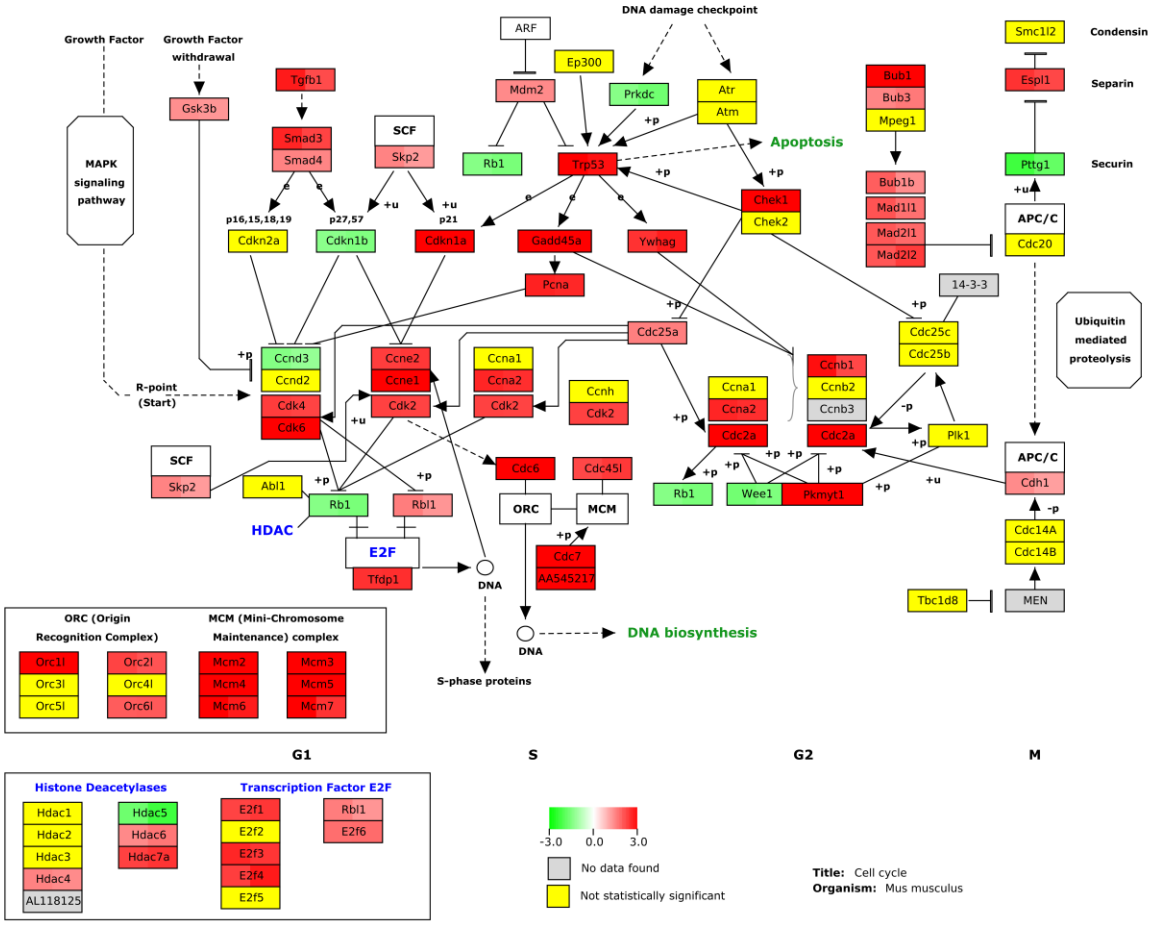
Supp. Figure S10: Gene count of a cluster of enriched processes that contain up-regulated genes in a *Gdf15*-specific AKI response.



Supp. Figure S11: Gene count of a cluster of enriched processes that contain down-regulated genes in a *Gdf15*-specific AKI response.



Supp. Figure S12: Gene expression visualization of the AKI pathway published by Husi et al, 2013. Left and right fill color represent the fold change observed in the dataset for WT and GDF15-KO mice comparisons, respectively (red up- and green down- regulated). Genes not expressed in the dataset are shown in grey, and genes expressed but not differentially expressed in yellow.



Supp. Figure S13: Expression profile visualization of the cell cycle genes. Left and right fill colors represent the fold change observed in the dataset for WT and GDF15-KO mice comparisons, respectively (red up- and green down- regulated). Genes not expressed in the dataset are shown in grey, and genes expressed but not differentially expressed in yellow.

Chapter 4

4. Sequencing of RNA from formalin-fixed paraffin-embedded laser-captured micro-dissected glomeruli

Daniel M. Borràs[1,2,3], Maria D. Sanchez-Niño[4], Alberto Ortiz[4], Joost P. Schanstra[2,3], Bart Janssen[1]

[1]*GenomeScan B.V, Plesmanlaan 1d, 2333BZ Leiden, The Netherlands.*

[2]*Institut National de la Santé et de la Recherche Médicale (INSERM), Institut of Cardiovascular and Metabolic Disease, Toulouse, France.*

[3]*Université Toulouse III Paul-Sabatier, Toulouse, France.*

[4]*Instituto Investigación Sanitaria - Fundación Jiménez Díaz - Universidad Autónoma de Madrid and Fundación Renal Íñigo Álvarez de Toledo - Instituto Reina Sofía de Investigación Nefrológica, Madrid, Spain.*

4.1. Introduction

Analysis of renal biopsies combined with blood and urine parameters can be very informative for patients suffering from chronic kidney diseases (CKD). It can be used to determine disease stages and progression rates (Haider et al., 2014; Wouters et al., 2015). Further research is required to unravel risk factors for CKD progression (Wouters et al., 2015). Systems-wide and systems biology approaches can be beneficial to further evaluate the impact of different models and kidney damage outcomes (Wouters et al., 2015; Eikrem et al., 2016). Even though fresh frozen biopsies would be an optimal source to perform this task, the availability of fresh-frozen tissue is limited compared to formalin-fixed, paraffin-embedded (FFPE) renal biopsies (Coudry et al., 2007; Hodgin et al., 2010; Eikrem et al., 2016). Biopsies are commonly fixed after extraction and further stored for sectioning and diagnosis of glomerular damage type. The strength of storing biopsies in FFPE is that these can be easily stored for long periods of time, allowing for large archives. The analysis of these archives would allow the characterization of large patient cohorts with many years of follow-up data collection. However, isolation of RNA species from FFPE material is still challenging due to chemical alterations and important RNA fragmentation (Masuda et al., 1999). Nevertheless, successful advances on isolating RNA from FFPE tissue have been reported (Coudry et al., 2007; Hodgin et al., 2010; Eikrem et al., 2016), increasing the chances of success to analyze these archives. In addition, laser-capture micro-dissection (LCM) technologies showed to be a successful approach to separate and enrich for tissue parts (Isenberg et al., 1976; Emmert-Buck et al., 1996; Coudry et al., 2007), allowing the isolation of RNA species from designated areas from any biopsy section, such as kidney glomeruli. Therefore, in this chapter we will assess the RNA sequencing of isolated total-RNA from LCMD glomeruli of FFPE biopsies.

Many studies (n=96[1]) have been reported using FFPE biopsies and microarrays for gene expression analysis. However, a lower number of studies (n=18[2]) were also performed with LCM in their methodology or by RNA-sequencing (n=15[3]). Only one[4] study was found that

[1] PubMed query accessed in 13-01-2017: (FFPE OR ("formalin-fixed" AND "paraffin-embedded")) AND "gene expression" AND (biopsy OR biopsies) AND microarray)
[2] PubMed query accessed in 13-01-2017: (FFPE OR ("formalin-fixed" AND "paraffin-embedded")) AND "gene expression" AND (biopsy or biopsies) AND (LCM or "laser-capture")
[3] PubMed query accessed in 13-01-2017: (FFPE OR ("formalin-fixed" AND "paraffin-embedded")) AND "gene expression" AND (biopsy OR biopsies) AND (RNA-seq OR "RNA sequencing")
[4] PubMed query accessed in 13-01-2017: (FFPE OR ("formalin-fixed" AND "paraffin-embedded")) AND "gene expression" AND (biopsy OR biopsies) AND (LCM OR "laser-capture") AND (RNA-seq OR "RNA sequencing")

used an RNA-seq approach for dissected tissue using LCM techniques to profile lung cancer progression (Morton et al., 2014). Because of the low number of related publications, there is a relevant need to develop or improve RNA-seq methodologies for analyzing FFPE kidney biopsies, as well as isolated glomeruli obtained by LCM. Here, we show the validation results of standard RNA sequencing using libraries obtained from FFPE tissue. In addition, we tested this approach for RNA isolated from glomerular LCM and FFPE samples.

4.2. Methods

4.2.1. Validation of library preparation and sequencing protocols for FFPE material

The performance of standard library preparation methods, such as NEBNext Ultra, for Illumina sequencing was evaluated for FFPE-tissue RNA. Total RNA was isolated at GenomeScan B.V. from 10 melanoma biopsies using Qiagen Allprep FFPE kit, and a mixture of RNA was prepared by equimolarly pooling. Dilution series from 200 ng to 1 ng were prepared to evaluate sequencing sensitivity. After rRNA depletion, cDNA libraries (NEBNext Ultra RNA-Seq kit) with dual indexing strategy were obtained. To compensate for the low input of lower dilution series, the number of PCR cycles was increased (Table 1). Library quality controls were performed to corroborate the expected library size (Fragment Analyzer) (Figure 1; Table 2). Sequencing was performed using a paired-end protocol on the Illumina Hiseq 2500 platform with SBS v4 reagents.

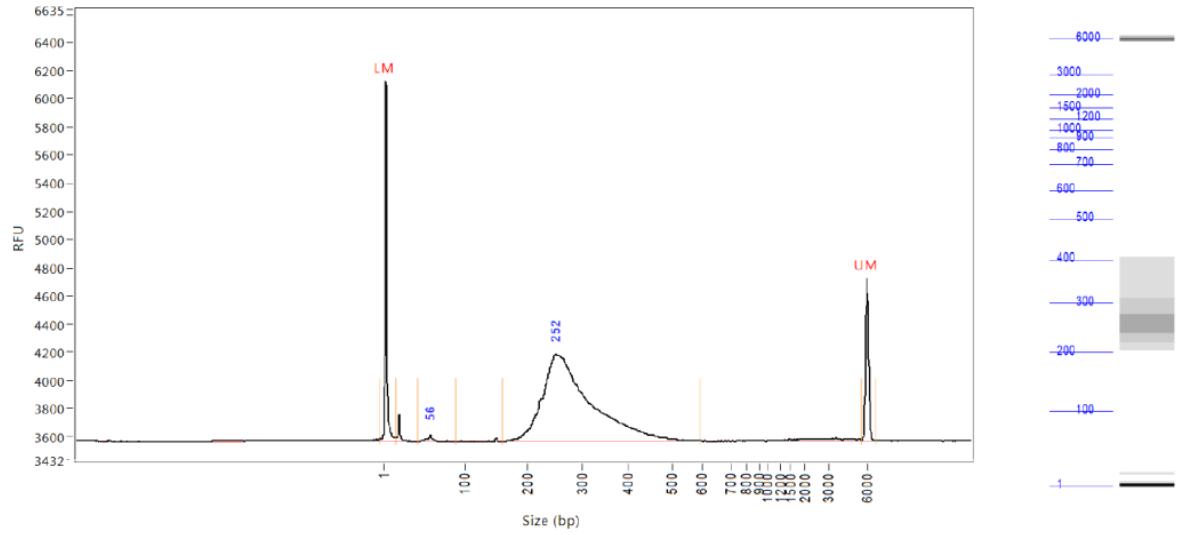


Figure 1: Fragment Analyzer output of NEB RNA val rRNA dep FFPE 200. Due to already degraded RNA from FFPE and additional fragmentation this size differs from intact RNA.

Table 1: Concentration series phage Lambda using NebNext Ultra

Input	*200 ng*	*100 ng*	*50 ng*	*25 ng*	*10 ng*	*5 ng*	*2.5 ng*	*1 ng*
PCR cycles	8	9	10	11	12	13	14	15
Size selection	Y	Y	Y	N	N	N	N	N

Table 2: Quality control of library preparation metrics from Fragment Analyzer.

Sample	*Input RNA (ng)*	*ng/µl*	*nmol/l*	*Avg. size (bp)*
NEB RNA val rRNA dep FFPE 200	200	10.2	58.3	289
NEB RNA val rRNA dep FFPE 100	100	3.3	17.3	310
NEB RNA val rRNA dep FFPE 50	50	2.0	9.9	329
NEB RNA val rRNA dep FFPE 10	10	1.6	8.0	329
NEB RNA val rRNA dep FFPE 5	5	2.6	12.3	352
NEB RNA val rRNA dep FFPE 1	1	2.0	11.1	302

Obtained reads in Fast-Q format were trimmed and filtered to remove sequencing adapters. Filtered reads were aligned to the reference genome sequence GRCh37.65 using Tophat v1.4.1 with the "*--stranded-reverse*" option, and with Bowtie v0.12.7. Alignments in standard alignment map format (S/BAM) were analysed using the RNA-SeQC v1.1.8 toolkit to retrieve RNA sequencing quality control measures (Table 3). In addition, specificity of the RNA sequencing library was determined by measuring read counts of features such as introns, exons and intergenic regions with Qualimap v2.0.1. Then, sensitivity of the dilution series was assessed by gene counts correlation using R software for statistical computing. Gene counts were calculated with HTSeq-count v0.5.4p5 and loaded to DESeq v1.10.1 R package for library size normalization. Finally, correlation matrix of r^2 values were calculated by fitting a linear model of log2 transformed normalized counts.

Table 3: Number of rRNA reads, percentage rRNA reads, and number of detected genes in data set (FPKM > 1).

Sample	*Mapped reads*	*rRNA reads*	*% rRNA*	*Genes found*
NEB RNA val rRNA dep FFPE 200	28,328,248	543,250	1.9	20,041
NEB RNA val rRNA dep FFPE 100	18,629,113	201,774	1.1	20,000
NEB RNA val rRNA dep FFPE 50	22,031,412	117,542	0.5	21,134
NEB RNA val rRNA dep FFPE 10	19,872,319	206,189	1.0	17,317
NEB RNA val rRNA dep FFPE 5	18,123,532	152,570	0.8	16,444
NEB RNA val rRNA dep FFPE 1	15,215,216	622,437	4.1	9,111

4.2.2. Proof of principle of sequencing RNA from laser-capture micro-dissected glomeruli from kidney FFPE tissue

Assessment of RNA sequencing for LCM glomeruli from FFPE tissue was performed using 4 available nephrectomies (2 patients) or renal tumorectomies (2 patients) were provided by associated partner Fundación Jiménez Díaz (FJD) (Table 4). Uninvolved tissue parts from the

tumorectomies can be considered as unaffected and were used to represent normal glomeruli (Hodgin, et al., 2010). Therefore, these samples contain fibrotic glomeruli and normal glomeruli, isolated from nephrectomies. To mimic the nature of kidney biopsies, only a piece of 2-3mm thick and 10-20 mm long was separated.

Table 4: Selection of samples from kidney nephrectomies.

Sample ID	Age (years)	Gender	Serum creatinine (mg/dl)	eGFR (ml/min/1.73 m^2)	Obstruction	Type	Fibrosis
7045-01-01	58	Woman	0.6	>60	No	Nephrectomy	Non-Fibrotic
7045-01-02	67	Woman	0.9	>60	No	Nephrectomy	Non-Fibrotic
7045-01-03	72	Woman	6.7	Dialysis	No	Nephrectomy	Fibrotic
7045-01-04	74	Woman	1.32	41.8	No	Nephrectomy	Fibrotic

Samples were processed for sectioning and LCM at the FJD. Sections of 5-6 μm were mounted on normal glass slides to be processed for LCM following adapted protocols for deparaffinization and staining for FFPE tissue (Table 4) (Hodgin et al., 2010; Grover et al., 2012; Jackson et al., 2013). Deparaffinization was performed using xylene as it showed good results (Jackson, et al., 2013), and is recommended by Qiagen for their RNA isolation kits (Table 5). Cresyl violet staining was used to identify cellular types as it was shown to be less aggressive with RNA species for further RNA isolation (Grover et al., 2012). Then a Cresyl Violet Acetate staining procedure was adapted from Grover et al., (2012) (Table 5). Dissection of fibrotic and normal glomeruli was performed at the FJD using the inverted microscope (Zeiss) with UV-A laser and PALM capture system. Then, the isolation of RNA was performed using a commercially available kit from Qiagen (RNeasy FFPE) following the manufacturer's protocol. Finally, the sequencing library to be sequenced on an Illumina HiSeq 2500 platform, was prepared using the same NEBNext Ultra protocol used for the validation of the library preparation and sequencing protocols for FFPE material.

Read sequences in FastQ format for each of the 4 samples were trimmed to remove sequencing adapter using Trimmomatic v.0.30. Then, reads were aligned to the human reference sequence version GRCh38.p2 using TopHat v.2.0.14 with default parameters except for "*--library-type fr-firststrand*", and with Bowtie aligner v.2.1.0. The alignment outputs in S/BAM format were sorted and indexed using Samtools v.1.2. Alignment quality metrics were collected using Qualimap v.2.0.1, and RNA-SeQC v.1.1.8. RNA feature counting was performed using HTSeq-counts v.0.6.1p1 for stranded libraries against the annotated features of the human genome in GTF format for the reference version GRCh38.p2.

Table 5: LCM slide preparation protocols

Step	***Deparaffinization.***
1	Mount sections onto the membrane slides.
2	Dry the slides at 56˚C. (overnight recommended if tissue sections are mounted by floating in warm DEPC-treated water)
3	Bake at 60˚C for 1-10 minutes to finish melting paraffin (1-2 minute recommended after overnight drying)
4	Deparaffinize with xylene (minimum 2 minutes and twice; maximum 15 minutes in total if Membrane Slides 1.0 PEN are used)
5	Wash with 100% ethanol, 1-2 minutes, twice, to remove excess of xylene.
7	Dry the slides shortly, 1-2 minutes, by leaning them against an RNase-free support on a lint-free absorbent paper.
6	Storage point. After drying, slides can be stored at -80˚C in an appropriate sealed RNase-free slide box with silica gel.
	Staining solution*
1	Dissolve 4% w/v Crystal Violet Acetate in 70% ethanol.
2	Agitate the staining solution for several hours to overnight at room temperature.
3	Filter to remove unsolubilized dye powder.
	Slide staining*
1	Without thawing the slides, transfer them to 100% ethanol for 30 seconds.
2	95% ethanol for 30 seconds.
3	75% ethanol for 30 seconds.
4	70% ethanol for 30 seconds. (slightly agitate the slides to dissolve possible polar soluble used compounds)
5	Stain with a dilution of 4% cristal violet in 70% ethanol for 30 seconds.
6	70% ethanol for 30 seconds. (slightly agitate to help remove excess of staining dilution)
7	75% ethanol for 30 seconds.
8	95% ethanol for 30 seconds.
9	100% ethanol for 30 seconds.
10	Repeat the 100% ethanol for 30 seconds.
11	Dry the slides shortly, 1-2 minutes, by leaning them against an RNase-free support on a lint-free absorbent paper.
12	Storage point. Slides can be stored at -80˚C in an appropriate sealed RNase-free slide box with silica gel.

**Staining protocols adapted from (Grover, et al., 2012)*

4.3. Results

4.3.1. Validation of library preparation and sequencing protocols for FFPE material

RNA sequencing libraries obtained from FFPE tissue showed an average insert size was a bit lower than usually expected for RNA sequencing (400 bp), but within the acceptable value ranges (200 - 700 bp) (Table 2). Overall, the total number of reads sequenced, on average 38.7M, were reduced by 17%, on average during the filtering and trimming steps. On average 20M reads (63%) were uniquely mapped to the human reference genome sequence GRCh37.65 (Table 6). Samples with input material ≥ 50 ng showed the higher percentages of mapped reads, whereas samples with 1 ng as input material showed the lowest (Table 6). Presence of ribosomal RNA carry over in the libraries was below 2% in most of the samples and 4% for the lowest input material of 1 ng (Table 3). All samples however, show a similar

distribution of reads among exonic, intronic and intergenic regions, being 23%, 61% and 16% on average, respectively (Table 7).

Table 6: Amount of raw, filtered and mapped reads and percentage of reads mapping of filtered reads.

Sample	***Raw reads***	***Filtered***	***(%)***	***Mapped***	***(%)***
NEB RNA val rRNA dep FFPE 200	51,885,558	44,541,316	86	28,328,248	64
NEB RNA val rRNA dep FFPE 100	31,670,180	27,210,978	86	18,629,113	69
NEB RNA val rRNA dep FFPE 50	37,199,686	31,820,052	86	22,031,412	69
NEB RNA val rRNA dep FFPE 10	37,128,894	30,554,342	82	19,872,319	65
NEB RNA val rRNA dep FFPE 5	33,817,000	27,422,216	81	18,123,532	66
NEB RNA val rRNA dep FFPE 1	40,463,678	32,287,048	80	15,215,216	47
Average	*38,694,166*	*32,305,992*	*83*	*20,366,640*	*63*

Table 7: Distribution of reads within genome annotation features.

Sample	***exonic***	***(%)***	***Intronic***	***(%)***	***intergenic***	***(%)***
NEB RNA val rRNA dep FFPE 200	3,136,835	23%	7,722,838	56%	2,885,282	21%
NEB RNA val rRNA dep FFPE 100	2,975,700	23%	7,859,929	62%	1,882,487	15%
NEB RNA val rRNA dep FFPE 50	4,337,096	24%	11,957,265	65%	2,096,065	11%
NEB RNA val rRNA dep FFPE 10	3,479,145	23%	9,744,110	65%	1,832,929	12%
NEB RNA val rRNA dep FFPE 5	3,282,375	23%	9,181,771	65%	1,738,987	12%
NEB RNA val rRNA dep FFPE 1	859,298	22%	2,107,400	54%	971,243	25%
Average	*3,011,742*	*23%*	*8,095,552*	*61%*	*1,901,166*	*16%*

The number of genes detected, was over 20,000 for samples above 50 ng of input material (Table 3). However, decreasing input material detected lower number of genes until 1 ng which showed the lowest with only 9,111 genes detected (Table 3). The number of genes detected for each dilution of input material showed a correlation of over 0.93 r^2 for input higher than 50 ng (Table 8). However, lower amounts of input material show increasingly lower sensitivity up to 0.55 - 0.59 r^2 of the 1 ng dilution.

Table 8: FFPE sample correlation r^2 for each level of input concentration

Dilution of correlated samples:	***200***	***100***	***50***	***10***	***5***	***1***
NEB RNA val rRNA dep FFPE 200	1					
NEB RNA val rRNA dep FFPE 100	0.95	1				
NEB RNA val rRNA dep FFPE 50	0.93	0.93	1			
NEB RNA val rRNA dep FFPE 10	0.86	0.86	0.85	1		
NEB RNA val rRNA dep FFPE 5	0.84	0.83	0.82	0.78	1	
NEB RNA val rRNA dep FFPE 1	0.59	0.58	0.57	0.56	0.55	1

4.3.2. Proof of principle of sequencing RNA from laser-capture micro-dissected glomeruli from kidney FFPE tissue

Using LCM on kidney tissue sections for all 4 samples, we obtained an average of more than 45 (n=9 to 66) glomeruli cross-sections (Table 9). RNA quality obtained from the isolated RNA of glomeruli cross-sections was <2.1 RIN score, which produced an RNA sequencing

library with an expected average size of ~300 bp (287-343 bp) (Table 9). However, the overall signal was very low (Figure 2) when compared to the overall signal obtained from the validation of FFPE sequencing (Figure 1). In addition, samples 2, 3 and 4 showed an unknown/unexpected signal with three peaks around 212 bp (Figure 2).

Table 9: LCM stats, and QC for the RNA isolation and library preparation.

Sample ID	# cross-sections	pg / µl / #	µm²	N cells*	pg/µl	RIN	ng/µl	nM/l	Avg. Size
7045-01-01	66	35.12	1,309,391	10,054	2,318	2.1	0.7458	3.576	343
7045-01-02	42	15.48	429,324	3,424	650	1	3.414	17.716	317
7045-01-03	9	69.78	209,932	1,771	628	1	1.0624	6.863	255
7045-01-04	65	2.431	1,537,360	11,771	158	1	0.5177	2.965	287
		LCM			RNA QC		Library QC		

**Number of cells was estimated using cell diameter 13 µm.*

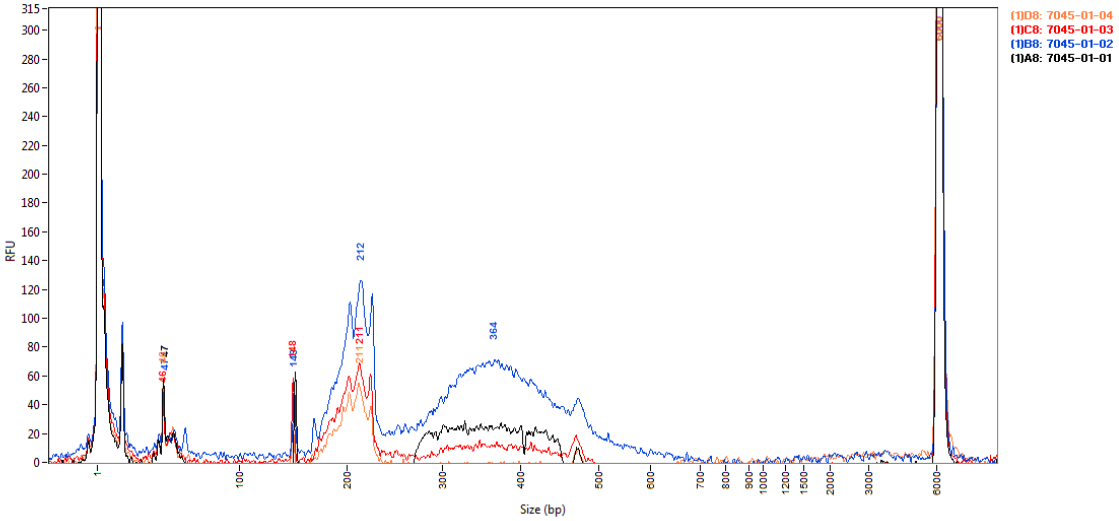


Figure 2: Fragment analyzer signal for RNA isolated from glomeruli cross-sections.

To evaluate the biological information retrieved from this RNA library preparation, we sequenced the libraries at low depth on a HiSeq2500 sequencer. We processed the reads as described above and extracted the gene/exon information by alignment against the human reference sequence. Overall, we obtained between 250,552 and 502,687 reads (Table 10). The maximum alignment percentage was 25% for sample 1, and lower for other samples (3-8%) (Table 10). Feature counting of mapped reads also shows low percentage of exonic reads, from ~1% to ~4% with sample 1 showing the highest percentage and sample 3 the lowest exonic percentage.

Table 10: Feature contribution of aligned low-depth sequenced reads.

Sample ID	*# reads*	*# aligned*		*# exonic*		*# no feature*	*No feature*	*% intronic*	*% intergenic*
7045-01-01	250,552	64,731	(25.8%)	1,044	(4.17%)	63,687	25.42%	13.04%	12.38%
7045-01-02	481,723	24,897	(5.2%)	403	(0.84%)	24,494	5.09%	2.78%	2.30%
7045-01-03	334,883	11,779	(3.5%)	234	(0.70%)	11,545	3.45%	1.83%	1.61%
7045-01-04	502,687	40,945	(8.1%)	798	(1.59%)	40,147	7.99%	4.57%	3.42%

4.4. Discussion

A common procedure applied to renal biopsies is to be FFPE after the extraction to facilitate the storage, and sectioning for the diagnosis of the glomerular damage type. This also contributes in increasing the FFPE archives that could be used for research purposes in detriment of the scarce availability of fresh-frozen tissue (Coudry et al., 2007; Hodgin et al., 2010; Eikrem et al., 2016). In this chapter, we evaluated the requirements for RNA sequencing of isolated material from FFPE tissue, to assess future possibilities of characterizing large patient cohorts with many years of follow-up data such as the cohorts potentially available from FFPE archives. It was known that the RNA isolated from FFPE material was the remnant of significant fragmentation and chemical alterations and a challenge for sequencing (Masuda et al., 1999). However, despite the limited quality of the RNA, we successfully quantified gene expression profiles from FFPE samples, clean of the interference of ribosomal RNA (<4%) (Table 3). These results show an increased sensitivity, when compared to similar work performed using microarray approaches (Hodgin et al., 2010; Mittempergher et al., 2011). We could observe the most probable effects from the degradation and chemical alterations of the FFPE procedures in the form of a reduction of the overall mapping observed (63% on average) (Table 6). In addition, the proportion of exonic mapped reads (23%) in conjunction with the proportions of intronic (61%) and intergenic (16%) mapped reads (Table 7), suggests a high proportion of carryover DNA material, most likely originating from the cross-linking in formalin between RNA and DNA molecules. The possibility of identifying interferences caused by degradation-associated FFPE effects may provide more accurate gene expression profiles than with other hybridization-based methods such as QPCR or microarrays. In addition, with the exposed methodology setup, we can suggest 50ng as the minimum input material required for an accurate quantification of expressed genes (>0.93 r^2) (Table 3; Table 8).

In renal diseases, the evaluation of glomerular lesions can be very informative to determine disease stages and progression rates (Haider et al., 2014; Wouters et al., 2015). The

enrichment and analysis of glomerular RNA showed to be a promising source of information to facilitate the diagnosis of glomerular lesions (Hodgin et al., 2010). By applying a similar approach as in the validation of RNA sequencing from FFPE material, we evaluated the possibility of profiling glomeruli's RNA expression using LCM to assess expression differences between fibrotic and normal glomeruli. The number of glomeruli cross-sections in the majority of samples was sufficient to obtain a good representation and confident gene expression profiles, while minimizing possible biases and tissue heterogeneity (>20 cross-sections; Table 4) (Hodgin & Cohen, 2010). Nevertheless, the sequencing information that could be extracted from these samples was almost null in terms of overall alignments and proportion of mapped exonic reads (Table 10). This might be due to the combination of low quantity and poor quality of the RNA that was isolated from the FFPE material (Table 9). Even sample one that showed the highest RNA integrity score (RIN 2.1) and concentration (2.3ng/µl) could not provide results comparable to previous work performed by Hodgin et al, 2010 using a similar approach with microarray technology. The low proportion of overall mapping (<25.8%) and exonic mapping (<4%) (Table 10), when compared to the results obtained for the validation of FFPE sequencing (63% and 23%, respectively), suggests that further work may be needed for accurately profiling the RNA expressed from LCM glomeruli, including extra tuning of the LCM and RNA isolation protocols.

In conclusion, we showed that gene expression profiling from FFPE samples is possible using RNA sequencing. However, additional work will be required to obtain reliable expression profiles from the low-quantity and low-quality RNA that is typically obtained when LCM is performed on glomeruli from FFPE kidney biopsies.

4.5. Acknowledgements

Arnoud Schmitz for his work on the FFPE-seq validation.

4.6. References

Coudry RA, Meireles SI, Stoyanova R, Cooper HS, Carpino A, Wang X, Engstrom PF, Clapper ML. 2007. Successful Application of Microarray Technology to Microdissected Formalin-Fixed, Paraffin-Embedded Tissue. J Mol Diagn JMD 9:70–79.

Eikrem O, Beisland C, Hjelle K, Flatberg A, Scherer A, Landolt L, Skogstrand T, Leh S, Beisvag V, Marti H-P. 2016. Transcriptome Sequencing (RNAseq) Enables Utilization of Formalin-Fixed, Paraffin-Embedded Biopsies with Clear Cell Renal Cell Carcinoma for Exploration of Disease Biology and Biomarker Development. PLoS ONE 11:e0149743.

Emmert-Buck MR, Bonner RF, Smith PD, Chuaqui RF, Zhuang Z, Goldstein SR, Weiss RA, Liotta LA. 1996. Laser capture microdissection. Science 274:998–1001.

Grover PK, Cummins AG, Price TJ, Roberts-Thomson IC, Hardingham JE. 2012. A simple, cost-effective and flexible method for processing of snap-frozen tissue to prepare large amounts of intact RNA using laser microdissection. Biochimie 94:2491–2497.

Haider DG, Masghati S, Goliasch G, Mouhieddine M, Wolzt M, Fuhrmann V, Hörl WH, Kaider A, Soleiman A. 2014. Kidney biopsy results versus clinical parameters on mortality and ESRD progression in 2687 patients with glomerulonephritis. Eur J Clin Invest 44:578–586.

Hodgin JB, Borczuk AC, Nasr SH, Markowitz GS, Nair V, Martini S, Eichinger F, Vining C, Berthier CC, Kretzler M, D'Agati VD. 2010. A Molecular Profile of Focal Segmental Glomerulosclerosis from Formalin-Fixed, Paraffin-Embedded Tissue. Am J Pathol 177:1674–1686.

Isenberg G, Bielser W, Meier-Ruge W, Remy E. 1976. Cell surgery by laser micro-dissection: A preparative method. J Microsc 107:19–24.

Jackson AF, Williams A, Moffat I, Phillips SL, Recio L, Waters MD, Lambert IB, Yauk CL. 2013. Preparation of archival formalin-fixed paraffin-embedded mouse liver samples for use with the Agilent gene expression microarray platform. J Pharmacol Toxicol Methods 68:260–268.

Masuda N, Ohnishi T, Kawamoto S, Monden M, Okubo K. 1999. Analysis of chemical modification of RNA from formalin-fixed samples and optimization of molecular biology applications for such samples. Nucleic Acids Res 27:4436–4443.

Mittempergher L, Ronde JJ de, Nieuwland M, Kerkhoven RM, Simon I, Rutgers EJT, Wessels LFA, Veer LJV. 2011. Gene Expression Profiles from Formalin Fixed Paraffin Embedded Breast Cancer Tissue Are Largely Comparable to Fresh Frozen Matched Tissue. PLOS ONE 6:e17163.

Morton ML, Bai X, Merry CR, Linden PA, Khalil AM, Leidner RS, Thompson CL. 2014. Identification of mRNAs and lincRNAs associated with lung cancer progression using next-generation RNA sequencing from laser micro-dissected archival FFPE tissue specimens. Lung Cancer Amst Neth 85:31–39.

Wouters OJ, O'Donoghue DJ, Ritchie J, Kanavos PG, Narva AS. 2015. Early chronic kidney disease: diagnosis, management and models of care. Nat Rev Nephrol 11:491–502.

Chapter 5

5. General Discussion

Parts of this chapter were adapted from Daniel M. Borràs[1] and Bart Janssen[1] - "The use of transcriptomics in clinical applications." (under revision).

[1]GenomeScan B.V., Pleasmanlaan 1d, Leiden

5.1. Future trends of sequencing technologies in diagnostics

In recent years, the number of research projects that included NGS as part of their approach has exponentially increased (Figure 1). The advancements in sequencing technologies and new approaches, as well as their increased accuracy and quality of data obtained facilitated this trend. A similar trend is also observed for those clinical publications that use NGS, however, with an expected delay in time (Figure 1). Finally, despite the very low numbers of publications reported that were using NGS in the field of kidney research, a significant proportion of them (76/178 in 2015) were oriented towards clinical research and diagnostics (Figure 1). Because of the continuous decrease of the sequencing costs observed in recent years, we hypothesize that the use of NGS for molecular testing will most likely keep increasing in the incoming years. Hence, the work presented in this thesis can aid in setting up the path towards a near future of NGS in clinical research.

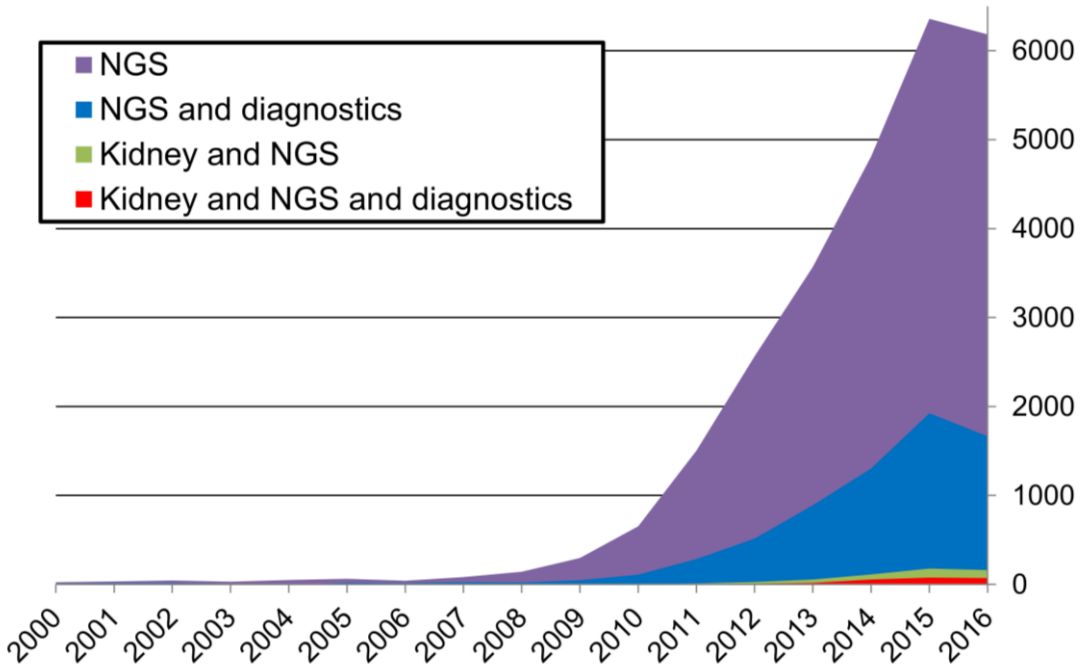


Figure 1: Number of yearly publications from 2000 to 2016 obtained using PubMed (16/03/2017) queries for: **NGS** *"(NGS OR (next generation sequencing)) NOT (microarray OR micro-array)*"; **NGS and diagnostics** "*(NGS OR (next generation sequencing)) NOT (microarray OR micro-array) AND diagnostic*"; **Kidney and NGS** "*(NGS OR (next generation sequencing)) NOT (microarray OR micro-array) AND (renal OR kidney OR nephrology)*"; and **Kidney and NGS and diagnostics** "*(NGS OR (next generation sequencing)) NOT (microarray OR micro-array) AND diagnostic AND (renal OR kidney OR nephrology)*". In 2015, the total number of publications reported in the field of kidney diagnostics was 19,843, only 178 used NGS in kidney research, and 76 used NGS in kidney research oriented to diagnostics.

5.2. Long-read sequencing approaches for improved resolution and characterization of complex regions for diagnostics

The advances of sequencing technologies, particularly in the direction of longer and more accurate reads, manifested the existing technical limitations of short-read sequencing, showing that for the resolution of complex regions in the reference human genome the required read length was estimated to be over 2000 bps (Lee and Schatz, 2012). In the case of *PKD1*, showcased in this thesis, current sequencing technologies used for the molecular testing in this complex gene include Sanger sequencing or newly developed short-read NGS approaches. Unfortunately, neither Sanger nor short-read NGS fulfil this >2000bps read length criterium. Hence, their power of resolution to resolve complex regions - which is the case for *PKD1* gene - is insufficient as discussed in Chapter 2. The use of long-read

Table 1: Medically relevant genes with inaccessible coding regions (problematic for short-read sequencing)*.

	Target region	# individuals	Known genes *(RefSeq, 22364 genes)*	Medically relevant genes *(ClearSeq IDP**, 3317 genes)*
Not properly covered *(Depth* ≤8)*	Genome wide	100%	1,404 (6.3 %)	63 (1.9 %)
		≥22%	5,375 (24 %)	362 (11 %)
	Exonic	100%	412 (1.8 %)	52 (1.6 %)
		≥22%	2,069 (9.3 %)	307 (9.3 %)

**Genome wide number of genes with poorly covered regions obtained from a published nine individual dataset by Sun et al., Human Mutation 2015.*

***ClearSeq Inherited Disease gene panel (Agilent Catalog, SureSelect DNA, design ID S0684402).*

approaches showed to be of value in many research studies when applied to resolve complex genetic situations, either previously known, or unknown, such as (1.) the reconstruction, resolution, and re-assembly of complex regions of the reference human genome including the closure of gaps (Chaisson et al., 2014; Huddleston et al., 2014; Steinberg et al., 2014); (2.) the characterization of entire gene families of complex genes for non-model or uncharacterized genomes (Larsen et al., 2014); (3.) the development of functional microsatellite markers (Grohme et al., 2013; Wei et al., 2014); (4.) the resolution of the highly repetitive, and complex central exon of *MUC5AC*, previously reported as an unknown region, or gap (Guo et al., 2014); (5.) the characterization and full resolution of the 100% GC-rich repeat region, and the mutation ranges of fragile X causing mutations in *FMR1* gene (Loomis et al., 2013). All these, are clear examples of the need for long-read sequencing data, which show minimal or very low reference bias, as well as low or no bias for high GC/AT regions and repetitive sequences.

Table 2: Medically relevant genes with repetitive elements potentially unresolvable by short-read sequencing.

		Known genes *(RefSeq, 22364 genes)*	**Medically relevant genes** *(ClearSeq IDP**, 3317 genes)*
Segmental Duplications	≥ 1 kbps	2,532 (11.3%)	699 (21%)
Simple Repeats	≥ 0.1 kbps	1,428 (6.4%)	462 (14%)
	≥ 1 kbps	360 (1.6%)	218 (6.60%)
Repeats*	≥ 0.1 kbps	2,290 (10.2%)	667 (20%)
	≥ 1 kbps	344 (1.5%)	270 (8.10%)

**Repetitive element RepeatMasker track.*
***ClearSeq Inherited Disease gene panel (Agilent Catalog, SureSelect DNA, design ID S0684402).*

The accessibility and resolvability of genomic regions, particularly for medically relevant genes, are of importance for the accurate development of diagnostic applications. Coding regions (exons) in the human genome that are not well represented (covered) when using short-read NGS data can be defined as inaccessible regions (Table 1) (Sun et al., 2015). Many inaccessible regions are inaccessible because of high GC content. Inaccessible regions tend to overlap with the so-called unresolvable regions. These, usually include repetitive regions and low complexity sequences of the reference genome such as segmental duplications (SD), simple repeats (SR) or other repetitive elements, as well as high GC/AT content regions and also gaps (Table 2) (Figure 2). The gaps in the human reference sequence are usually complex regions that have simply not been resolved and therefore remain unknown (Chaisson et al., 2014; Huddleston et al., 2014; Steinberg et al., 2014) (Guo et al., 2014). The combination of segmental duplications, high GC-content, and/or repetitive elements impedes unique alignments against the reference sequence. Therefore, affecting the sequencing depth of short-reads in these regions because of the low mapping quality (Figure 2). Within all these inaccessible and unresolvable coding regions, we find some disease-associated genes that are routinely screened in diagnostic procedures (Figure 3). If the gene of interest lies within an inaccessible or unresolvable region, the use of short-read sequencing approaches would reduce the chances of resolving genetic variants that may be potentially pathogenic.

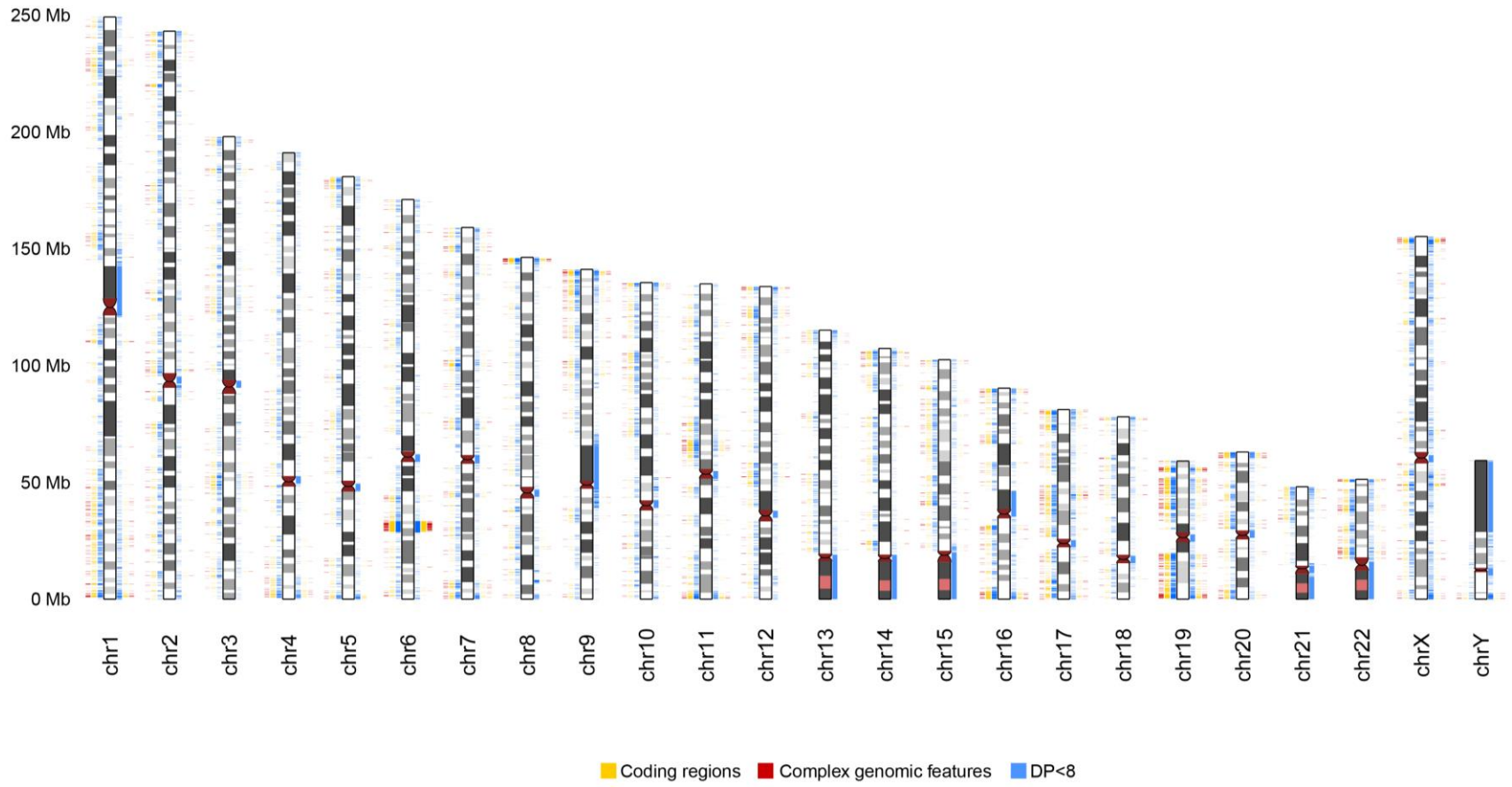


Figure 2: *Genomic overview of* inaccessible *regions* (blue), *from which the overlapping coding* (yellow) *and* unresolvable *regions* (red) *are highlighted*. Data shown in the left and right sides (top and bottom for the magnified chr16) correspond to published WES and WGS data, respectively (Sun et al., 2015).

Recently developed NGS-based (short-reads) diagnostic approaches for ADPKD showed 95% - 100% sensitivity and specificity when compared to standard diagnostics using Sanger combined with multiplex ligation-dependent probe amplification (MLPA) (Eisenberger et al., 2015; Trujillano et al., 2014; Tan et al., 2014). However, the comparison between two diagnostic approaches that both could suffer from lower resolution because of the use of short-reads does not demonstrate the true diagnostic performance of either of these approaches. When taking into consideration all patients included from a cohort, including the ones that could not be accurately genotyped, the diagnostic performance reported by newly developed short-read NGS methods could only provide a clear diagnosis for 115 out of 183 (Rossetti et al., 2012), 16 out of 25 (Tan et al., 2014), 10 out of 12 (Trujillano et al., 2014), 35 out of 55 (Eisenberger et al., 2015), and 24 out of 28 (Mallawaarachchi et al., 2016) screened ADPKD patients. The methodological sensitivity and specificity reported was >97% (97% - 100%) when comparing the NGS-based results with the results of the standard Sanger-based diagnostics (Rossetti et al., 2012; Tan et al., 2014; Trujillano et al., 2014; Eisenberger et al., 2015; Mallawaarachchi et al., 2016). This high sensitivity and specificity suggests that short-read NGS-based ADPKD diagnostics approaches are strongly comparable to Sanger-based ADPKD diagnostics. However, it also shows that both approaches have the same technical limitations, namely short reads, and were not fully able to overcome the high complexities of the *PKD1* gene loci.

An enhanced resolution of complex regions of the reference human genome allowed the closure of gaps, and the reconstruction of low complexity sequences that was only possible using long-read sequencing (Chaisson et al., 2014; Huddleston et al., 2014; Steinberg et al., 2014). Reduced mappability of NGS short-reads interfere with the variant calling algorithms, producing undesired false positives and negative calls (Lee and Schatz, 2012). The results shown in Chapter 2 about *PKD1* diagnostics led to our proposal of a long-range PCR (LR-PCR) targeted approach to sequence long-read amplicons for *PKD1* and *PKD2* genes using SMRT sequencing approach with the PacBio RSII platform. The proposed pipeline is still at early stages of development and will require to be further validated with a larger cohort of ADPKD patients. However, we showcased that a reliable diagnostics approach that can resolve complex genetic setups is required, indispensable, and possible for clinical applications. We believe that this is the very beginning of a new era of long-read diagnostics for complex genetic regions. In addition, long-read approaches may probably also replace

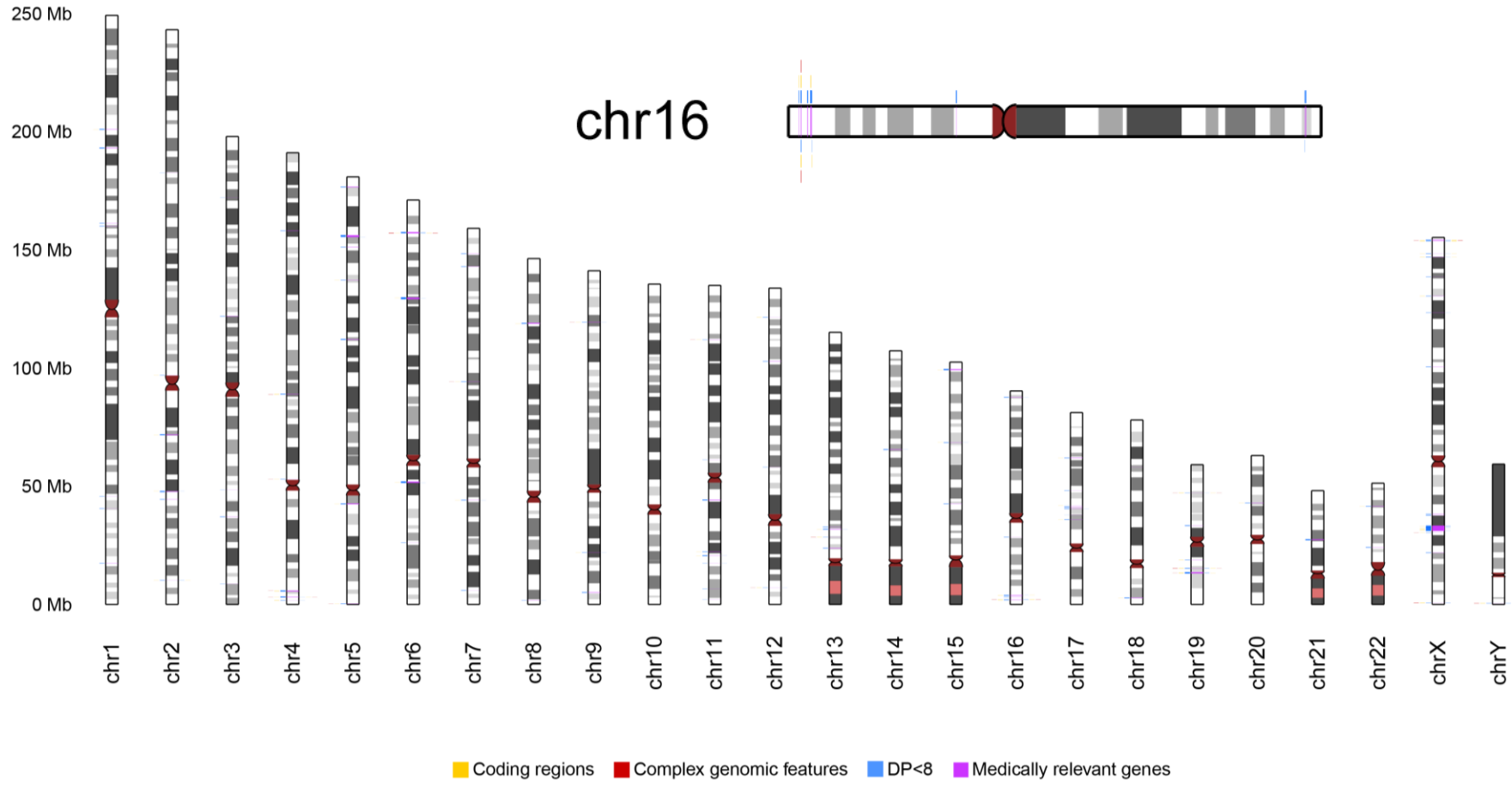


Figure 3: Medically relevant genes (pink) with inaccessible (blue) and unresolvable (red) coding regions (yellow). Data shown on the left and right sides (top and bottom for the magnified chr16) correspond to published WES and WGS data, respectively (Sun et al., 2015).

some other established diagnostic approaches simply because of the increased value for haplotyping, identifying large insertion/deletion breakpoints, as well as the low or no reference-biased variant calls that the majority of short-read approaches cannot currently offer. Some examples of the benefits of long-read sequencing for medically-relevant complex genes include: (*1.*) human leucocyte antigen (HLA) genotyping (Mayor et al., 2015); (*2.*) *FMR1* expanded CGG-repeat haplotyping (Loomis et al., 2013); (*3.*) genotyping of *MUC5AC* central exonic gap, including SNVs, SNPs, CNVs, and SDs (Guo et al., 2013).

5.3. RNA sequencing and system approaches for the identification and screening of disease biomarkers in complex diseases

Omics approaches when applied into clinical research and diagnostics, despite being quite costly, may be used to replace multiple tests of individual biomarkers. These individual biomarker tests would be required in increased numbers if these global approaches would not be available. Wide or global approaches such as RNA-Seq can screen for multiple disease determinants in a single run and obtain information on the cause of the disease as well as on the potential response to treatments among other factors. However, these wider screening approaches are still not commonly preferred over single gene PCR tests. The use of PCR for monitoring one or several genes is definitely cheaper than a complete expression profiling in the terms of price per sample. However, when considering the price per gene screened and the quality and quantity of information obtained, RNA-Seq offers a significant advantage. For instance, a 1 tier Fluidigm (multiple reaction RT-qPCR based platform) assessing 90 transcripts would cost 22 Euro (25$) per sample or 0.24 Euro (0.27$) per gene (adapted from Chaussabel 2015), while an mRNA-Seq would cost under 400 Euro per sample but less than 0.04 Euro per gene, considering that RNA-Seq potentially yields thousands of genes (Chapter 1, Table 8). In addition, RT-qPCR-based measurements will provide only the signal referent to one single exon or region while RNA-seq would provide signal for the entire gene structure including splicing isoforms and presence of exon variants. Reducing the complexity of disease diagnostics towards a few key genes may reduce the immediate costs of the clinical assay. However, the cost of NGS is and will keep decreasing with the advances and developments of new sequencing chemistries and platforms, making NGS approaches even more affordable and accessible in the future, e.g. the latest high-throughput sequencer NovaSeq 6000 announced by Illumina.

RNA-Seq approaches have the potential to be widely applied in clinical research and molecular testing since they are reliable and reproducible (Chaussabel, 2015). Given the mass of available knowledge about the human genome and disease biomarkers, sequencing approaches should be considered as a potential option for the future development of clinical research and diagnostic assays. Some examples of the evolution of clinical research based on NGS approaches show their potential towards the development of personalized medicine. Renkema *et al.*, (2014) reviewed several applications for chronic kidney disease (CKD) highlighting that NGS-based DNA genetic testing can reduce the costs and turnaround time in diagnostics of steroid resistant nephritic syndrome and autosomal dominant polycystic kidney disease (Renkema et al., 2014). The authors emphasize on the need of WGS approaches to identify nephronophthisis causative genetic variants due to oligogenic inheritance for genes such as *NPHP2*, *NPHP3*, and *AHI1*. In addition, NGS approaches allowed researchers to achieve a greater understanding of diseases and the pathogenesis of genetic disorders. Systems biology approaches that integrate data from various sources could also be used to screen for potential drug targets, for instance ADPKD animal models were used as an example to determine efficacy of vasopressin receptor V2 antagonist (Renkema et al., 2014). Using also a system wide unbiased RNA-seq approach, we showcased in Chapter 3 that the analysis of expression profiles of AKI mice kidneys could be very informative to characterize one of the three most common causes of AKI observed in humans, which include ischemia, toxic exposure and obstruction (Yang et al., 2010). Overall, given the latest tendency of clinical research towards the advancement of personalized medicine approaches, we may predict an increase of research projects to develop system-wide and omics-based clinical applications. Several studies were recently reviewed by Chaussabel., (2015) to provide a detailed vision of the perspectives of some omics approaches, in particular for blood transcriptomics, in clinical research and diagnostics. Applications reviewed included neurological disorders (Autism and Alzheimer), organ transplant rejection and risks signature genes (liver, heart, kidney and bone marrow), and a wide variety of different affections such as exposure of environmental factors, respiratory diseases, allergy, stroke, infections, and diabetes among others. The use of system-wide and omics approaches such as blood transcriptomics showed to be beneficial in clinical research in terms of types of diseases that can be characterized, diagnostic tests developed, and potential therapeutic targets identified. In Chapter 3 we discussed the example of an unbiased RNA-seq approach to characterize renal diseases such as AKI. The results obtained using this approach could identify gene

trends and pathways activated independently for each different condition (Chapter 3, Table 5). The transcription factor signature identified could be considered for further research as a potential AKI driver, and become key to the identification of AKI therapeutic targets (Chapter 3, Table 3; Chapter 3, Table 4).

Other global approaches such as the current versions of genome wide gene expression microarrays, despite an improved and reasonably accurate probe design, still rely on the hybridization of fluorescent DNA to *"quantify"* expression. DNA hybridization methods are known to produce some background noise that can interfere with the true signal. There have been several studies and reviews that address this topic while comparing different technologies in terms of throughput, accuracy, cost and efficiency. For instance, Marioni *et al.,* (2008) assessed the technical differences between micro-array technology and RNA-Seq. Despite the decent high throughput set-up that micro-arrays offer, the authors highlight the implicit high background noise of micro-arrays due to cross-hybridization. The methodology for controlling this noise in addition to the differences in the design of the hybridization probes makes micro-array results almost impossible to combine with other types of experiments (Marioni et al., 2008). This effect is mitigated in NGS approaches which showed much higher resolution, fewer artifacts, greater coverage, and a wider dynamic range than micro-arrays (Park, 2009). These factors as well as many other characteristics of microarray technology and RNA-Seq have been extensively compared and reviewed since the release of RNA-Seq in 2009 (Marioni et al., 2008; Mimura et al., 2014; Wang et al., 2009). It must be emphasized that the detection of low abundance transcripts could only be accurately performed by using RNA-Seq technology (Brennan et al., 2012). This, as well as the ability to sequence unknown regions makes sequencing approaches more sensitive and complete than micro-arrays, detecting many more differentially expressed genes (>25% more) in the case of RNA-seq (Sultan et al., 2008). Since RNA-seq gene quantification is based on the sequence alignments and not on hybridization fluorescence signal, the analysis of gene construct models with RNA-seq, such as Gdf15-KO mice analyzed in Chapter 3, can provide much increased precision and validation of the functionality of the Gdf15-KO construct (Chapter 3, Figure 1; Chapter 3, Supp. Figure 1). Our results show that the expression of aberrant transcripts, such as in the Gdf15-KO, would have passed completely unnoticed with a hybridization-based expression profile, such as microarrays, most likely leading to undesired false results or conclusions. Microarrays are still widely used in clinical research and diagnostics, and will most likely still have a complementary role in wider systems- or

omics-based applications (Schumacher et al., 2015). We showed that the use of global omics approaches such as RNA-seq showed increased potential to characterize complex diseases such as AKI. In particular, if the transcripts analyzed express different isoforms and haplotypes, and even more if the sequences of these transcripts are not well annotated or simply unknown. The development of future clinical research and diagnostic tests will certainly benefit from the quantity and quality of the information obtained using RNA-seq approaches in a single run. The molecular testing of AML can be considered a clear example of the future of RNA-seq in clinical diagnostics (Griffioen et al., 2016).

Global omics-based NGS procedures have sufficient throughput and sensitivity to identify system-wide DNA modifications, even those ones that were not intended to be screened for. These unsolicited findings opened some discussions about the ethical and practical aspects of reporting these analysis results. If system-wide approaches are to be developed and implemented in the diagnostic field, the information obtained, and provided to the patient would require some previous filtering. However, the distinction between what are relevant intended findings and relevant but unsolicited findings should be drawn by the patient with an educated consent. We believe that information and transparency is crucial to provide the necessary environment for decision making. In the end, adequately informed patient that knows about the analysis test performed and the results that could be obtained will consent to receive the level of information requested (Bijlsma et al., 2016).

5.4. Sample and data archives for improved clinical research

Sample availability, is usually limited in the case of human tissue, and much more for fresh-frozen tissue samples. This makes the process of sample selection a key step for the correct understanding of tissue associated biological processes. Blood samples, among other biofluids such as saliva, are one of the easiest tissue samples to obtain (Devonshire et al., 2013). Many studies have been performed with blood samples to detect biomarkers and molecular determinants. However, the differentially expressed transcripts detected in tissue may be underrepresented in blood or only detectable in later disease stages (Chen and Snyder, 2013). Despite these drawbacks, there are successful studies that characterized blood samples using transcriptomics, for instance in immunological diseases (Chaussabel, 2015). Other types of biofluids may also be easily obtainable making them an interesting target source for many researchers that want to develop new diagnostics tools using these non-invasive tissue sources such as saliva or urine (Devonshire et al., 2013; Suthanthiran et al.,

2013). Most tissue sample types are either not that easily accessible such as heart, liver and other internal organ samples, or completely unavailable such as brain tissue. If sample availability is scarce, it makes the switch towards global analysis or omics approaches even more necessary since all information, such as genes, variants or proteins, expressed in such samples would be measured at once. If this approach were to be applied as routine diagnostic tests, it would serve two purposes: (*1.*) to perform the required diagnostics for the patient at the moment of the test; (*2.*) provide a wider overview of the patient status that could be stored in a database and used for future research, if the patient provided explicit consent. It is common practice for tissue biopsies to be formalin-fixed and paraffin-embedded (FFPE) to maintain tissue structure for sectioning, and for long term storage. This provides large amounts of samples in FFPE archives that could be used for research purposes and compensate for the scarce availability of fresh-frozen tissue. The possibility of analyzing these large FFPE sample archives, with many years of follow up data, would facilitate larger clinical studies with improved statistical power (Chen and Snyder, 2013). The use of omics approaches to analyze large FFPE archives would make a difference for those complex diseases where sample numbers are limited, as well as providing the time required for the collection of rare disease samples. However, as discussed in Chapter 4, the sequencing of FFPE samples is still challenging.

When the total RNA is extracted from a biological sample we must take into consideration that its cell composition is an important source of variability. Tissue samples, for instance, contain different cell types that will express their own gene repertoire. For RNA-seq, the sample mRNA composition may also be influenced by internal, and external factors (i.e. nutrition, circadian stage, cellular cycle, stress, exercise or disease state). Sasagawa *et al.*, showed that single cell transcriptome analysis using RNA-Seq was able to identify and quantify non-genetic cellular heterogeneity, and even differentiate cell types and cell cycle phases of a single cell type (Sasagawa et al., 2013). Therefore, it is important to use appropriate methods for targeted cell type enrichment, if possible, such as laser capture micro dissection (LCM) for selecting tissue areas from tissue slides, cell sorting for enriching the cell fraction of interest, or centrifugation for separating the desired cell population (Todd and Kuo, 2002; Taussig et al., 2010; Devonshire et al., 2013; Sasagawa et al., 2013; Gutierrez-Arcelus et al., 2015). the results presented in Chapter 4, showed that human kidney biopsies may have too few glomeruli. Nevertheless, these results showed that it is possible to obtain glomeruli specific expression profiles using RNA-seq from LCM glomeruli of FFPE tissue

sections. Further advances in the development of a standardized protocol are still required to make the transcriptome profiles of the FFPE tissue archives fully available for clinical research.

Other global approaches, such as hybridization-based methods including microarrays and PCR, are bound to existing (and potentially limited or biased) knowledge about the transcriptome, genome, and known variant annotations, and their association to possible disease phenotypes. Advances in the understanding of the human genome, function of genes and their link with disease phenotypes will lead to improvements in the design of PCR tests and micro-array experiments. However, currently only a few organisms such as *Homo sapiens* have a well-studied transcriptome and our understanding of its complexity is still far from being complete (Marioni et al., 2008). Therefore, tests performed to analyze more complex and heterogenic diseases are more challenging to implement. The use of whole genome, transcriptome or proteome approaches, including RNA-Seq, would facilitate the retrieval of relevant data of complex or rare diseases if these approaches were used in a larger scale, for instance with FFPE sample archives. This, coupled with the increasing number of databases for storage and easy access to data, would provide the basis for larger and more complete studies with data that maintain scientific relevance for many years. The switch to global approaches such as whole transcriptome analysis will enable better prediction of disease onset, outcome, severity, treatment response and in general easier patient management (Chaussabel, 2015). However, despite the numerous advantages that system-wide approaches may offer to clinical research, many of these approaches are yet to be widely implemented in a diagnostic setting. This may be because of the general understanding that sequencing-based approaches such as RNA-seq has a cost-benefit risk when considered majorly from a monetary but not a medical point of view. If everything that needs to be measured is indeed measured by a transcriptome assay, there would be no market for dedicated assays, kits, and instruments for different tests and diagnostics (Chaussabel, 2015). Moreover, the benefits of the shift to RNA-sequencing would potentially improve clinical research and disease diagnostics, and reduce health care burden especially for complex diseases. In this hypothetic situation, diagnostics for autoimmunity, cancer, cardiovascular diseases, infectious diseases, neurological diseases, nutrition deficiencies, pregnancy tests, as well as disease severity, onset, outcome and response to treatments could be monitored from a single centralized laboratory (Berry et al., 2010; Fehlbaum-Beurdeley et

al., 2012; Chao et al., 2013; Sarwal and Sigdel, 2013; Chaussabel, 2015). Hence, approaching the health care system to the so called personalized medicine.

The significant progresses in high throughput technologies such as genomics, transcriptomics, proteomics and peptidomics among others, leads to the personalized high-throughput precision medicine. The traditional symptom-oriented diagnosis and treatments would be complemented with individual molecular profiles of the patients, allowing a much better and efficient treatment (Chen and Snyder, 2013). In this situation, sequencing as a continuously improving high throughput precision omics technology would greatly facilitate this process. The detailed information of a transcriptome profile or exome profile would reflect potential physiological changes at the sample collection time. Additionally, the integration with other omics would enhance the scientific research process. Eventually, this would translate into improved health care by monitoring the patient health status and by applying personalized and preventive treatments. At first, one may think that this would increase the costs of the health care system since some of these omics assays are currently quite expensive. However, we argue that this would especially help the characterization and understanding of complex diseases (Huang and Mucke, 2012; Shah et al., 2012; Codina-Solà et al., 2015) and, in the long term, globally reduce the health care burden (Chen and Snyder, 2013).

If global screening approaches were to be routinely implemented in clinical applications the quantity of available data would increase the need of data handling measures. High throughput applications can produce a tremendous amount of data, which is challenging to process, handle and properly annotate with the purpose of further clinical interpretation. This calls for data integration in centralized databases to facilitate the analysis and interpretation of the collected results. This would stimulate advances in computer technologies and databases as well as bioinformatics that would improve and continuously increase the available knowledge. In the end, the objective of storing all these data and clinical results is to make it easier for researchers and clinicians to mine these databases with the appropriate algorithms to make more accurate and elaborated medical decisions. For this, comprehensive databases are required which would store health records, variant calls, expression profiles and all other patient related molecular information (Chen and Snyder, 2013). There are currently many databases that can provide comprehensive functional annotations such as The Catalogue of Somatic Mutations in Cancer (COSMIC) (Forbes et al., 2015) which is a comprehensive collection of somatic mutations for human cancer, or the Leiden Open

(source) Variation Database (LOVD) (Fokkema et al., 2011) which provides a tool to collect and display DNA variants. These databases may potentially facilitate the complex process of annotation of high throughput sequencing data. It is through the small effort of global sharing of particular findings that these databases are greatly improving over the years in quantity and quality of annotations. Other currently ongoing efforts are thrown into database collections of gene expression datasets such as the Expression Atlas (Petryszak et al., 2014). In this case gene expression profiles are publicly available and accessible and provides information about different organisms, expression patterns and biological conditions among others. The Expression Atlas includes RNA-Seq experiment data as well as microarray experiment data which can be re-analyzed through a web portal (www.ebi.ac.uk/gxa). There are other similar initiatives focusing on kidney research such as Nephroseq and the Renal Gene Expression Database (Zhang *et al*., 2014). Please refer to the available online resources for additional information (www.nephroseq.org; rged.wall-eva.net). Additionally, there is also the possibility of using publicly available RNA-Seq and microarray datasets in combination with clinical data. This is the case of the Gene Expression Omnibus (GEO) from NCBI (Barrett et al., 2013), or the European Genome-Phenome Archive (EGA) (Lappalainen et al., 2015), which offer international repositories of microarray and NGS datasets submitted by the research community. With these initiatives, the research community would certainly benefit from the work of any other researcher and would increase the statistical power of the studies, and the accuracy of diagnostics applications.

5.5. Summary

The increasing number of sequencing approaches and methodologies being developed and currently available will facilitate future clinical research of new diagnostic applications. The identification of strengths and limitations of new approaches, as well as the research question to answer is crucial to find the appropriate method in a case by case basis. In this thesis, we presented a selection of three renal diseases that could benefit from the use of new sequencing technology approaches.

In Chapter 2, we have shown that the diagnosis of ADPKD, particularly for the complex *PKD1* gene, can certainly benefit from the use of single-molecule long-read sequencing. In addition, we discussed in Chapter 5.2 the challenges of resolving complex genomic regions such as *PKD1*, and the extent of these regions into other medically relevant genes that would also benefit from a similar long-read based approach.

In Chapter 3, using RNA-seq we could characterize the expression profiles associated with AKI, as well as the expression profiles of AKI with Gdf15 deficiency. We showed that an unbiased RNA-seq approach can identify response-associated AKI drivers such as transcription factors. Furthermore, we discussed in Chapter 5.3 the strengths of RNA-seq and systems-based approaches over other and "cheaper" methodologies, and its increased value in the future of clinical research and diagnostics.

In Chapter 4, we provided proof of principle that the sequencing of RNA from LCM isolated glomeruli for FFPE archived samples is plausible. In addition, we elaborated in Chapter 5 the impact of the accessibility of the expression profiles of FFPE sample archives in the future of clinical research.

Overall, we showed that sequencing approaches offer new possibilities in the research field of renal diseases, and discussed their increased value in the future of clinical research and diagnostics.

5.6. References

Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A, Lee H, et al. 2013. NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res 41:D991–D995.

Berry MPR, Graham CM, McNab FW, Xu Z, Bloch SAA, Oni T, Wilkinson KA, Banchereau R, Skinner J, Wilkinson RJ, Quinn C, Blankenship D, et al. 2010. An interferon-inducible neutrophil-driven blood transcriptional signature in human tuberculosis. Nature 466:973–977.

Bijlsma RM, Bredenoord AL, Gadellaa-Hooijdonk CG, Lolkema MP, Sleijfer S, Voest EE, Ausems MG, Steeghs N. 2016. Unsolicited findings of next-generation sequencing for tumor analysis within a Dutch consortium: clinical daily practice reconsidered. Eur J Hum Genet 24:1496–1500.

Brennan EP, Morine MJ, Walsh DW, Roxburgh SA, Lindenmeyer MT, Brazil DP, Gaora PÓ, Roche HM, Sadlier DM, Cohen CD, Godson C, Martin F. 2012. Next-Generation Sequencing Identifies TGF-β1-Associated Gene Expression Profiles in Renal Epithelial Cells Reiterated in Human Diabetic Nephropathy. Biochim Biophys Acta 1822:589–599.

Chao S, Ying J, Liew G, Marshall W, Liew C-C, Burakoff R. 2013. Blood RNA biomarker panel detects both left-and right-sided colorectal neoplasms: a case-control study. J Exp Clin Cancer Res 32:44.

Chaussabel D. 2015. Assessment of immune status using blood transcriptomics and potential implications for global health. Semin Immunol 27:58–66.

Chen R, Snyder M. 2013. Promise of personalized omics to precision medicine. Wiley Interdiscip Rev Syst Biol Med 5:73–82.

Codina-Solà M, Rodríguez-Santiago B, Homs A, Santoyo J, Rigau M, Aznar-Laín G, Campo M del, Gener B, Gabau E, Botella MP, Gutiérrez-Arumí A, Antiñolo G, et al. 2015.

Integrated analysis of whole-exome sequencing and transcriptome profiling in males with autism spectrum disorders. Mol Autism 6:21.

Devonshire AS, Sanders R, Wilkes TM, Taylor MS, Foy CA, Huggett JF. 2013. Application of next generation qPCR and sequencing platforms to mRNA biomarker analysis. Methods 59:89–100.

Eisenberger T, Decker C, Hiersche M, Hamann RC, Decker E, Neuber S, Frank V, Bolz HJ, Fehrenbach H, Pape L, Toenshoff B, Mache C, et al. 2015. An Efficient and Comprehensive Strategy for Genetic Diagnostics of Polycystic Kidney Disease. PLOS ONE 10:e0116680.

Fehlbaum-Beurdeley P, Sol O, Désiré L, Touchon J, Dantoine T, Vercelletto M, Gabelle A, Jarrige A-C, Haddad R, Lemarié JC, others. 2012. Validation of AclarusDx™, a blood-based transcriptomic signature for the diagnosis of Alzheimer's disease. J Alzheimers Dis 32:169.

Fokkema IFAC, Taschner PEM, Schaafsma GCP, Celli J, Laros JFJ, Dunnen JT den. 2011. LOVD v.2.0: the next generation in gene variant databases. Hum Mutat 32:557–563.

Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H, Ding M, Bamford S, Cole C, Ward S, Kok CY, Jia M, et al. 2015. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res 43:D805–D811.

Griffioen M, Arindrarto W, Borràs DM, Diessen SAME van, Meijden ED van der, Honders MW, Alloul M, Jedema I, Kroes WGM, Valk PJM, Janssen B, Bergen CAM van, et al. 2016. Whole Transcriptome Sequencing (RNAseq) As a Comprehensive, Cost-Efficient Diagnostic Tool for Acute Myeloid Leukemia. In: 58th Anual Meeting and Exposition of the Aerican Society of Hematology, San Diego.

Guo X, Zheng S, Dang H, Pace RG, Stonebraker JR, Jones CD, Boellmann F, Yuan G, Haridass P, Fedrigo O, Corcoran DL, Seibold MA, et al. 2013. Genome Reference and Sequence Variation in the Large Repetitive Central Exon of Human MUC5AC. Am J Respir Cell Mol Biol 50:223–232.

Gutierrez-Arcelus M, Ongen H, Lappalainen T, Montgomery SB, Buil A, Yurovsky A, Bryois J, Padioleau I, Romano L, Planchon A, others. 2015. Tissue-Specific Effects of Genetic and Epigenetic Variation on Gene Regulation and Splicing. PLoS Genet 11:e1004958.

Huang Y, Mucke L. 2012. Alzheimer Mechanisms and Therapeutic Strategies. Cell 148:1204–1222.

Lappalainen I, Almeida-King J, Kumanduri V, Senf A, Spalding JD, Ur-Rehman S, Saunders G, Kandasamy J, Caccamo M, Leinonen R, Vaughan B, Laurent T, et al. 2015. The European Genome-phenome Archive of human data consented for biomedical research. Nat Genet 47:692–695.

Loomis EW, Eid JS, Peluso P, Yin J, Hickey L, Rank D, McCalmon S, Hagerman RJ, Tassone F, Hagerman PJ. 2013. Sequencing the unsequenceable: Expanded CGG-repeat alleles of the fragile X gene. Genome Res 23:121–128.

Mallawaarachchi AC, Hort Y, Cowley MJ, McCabe MJ, Minoche A, Dinger ME, Shine J, Furlong TJ. 2016. Whole-genome sequencing overcomes pseudogene homology to diagnose autosomal dominant polycystic kidney disease. Eur J Hum Genet 24:1584–1590.

Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. 2008. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. Genome Res 18:1509–1517.

Mayor NP, Robinson J, McWhinnie AJM, Ranade S, Eng K, Midwinter W, Bultitude WP, Chin C-S, Bowman B, Marks P, Braund H, Madrigal JA, et al. 2015. HLA Typing for the Next Generation. PLoS ONE 10:.

Mimura I, Kanki Y, Kodama T, Nangaku M. 2014. Revolution of nephrology research by deep sequencing: ChIP-seq and RNA-seq. Kidney Int 85:31–38.

Nephromine.

Park PJ. 2009. ChIP-Seq: advantages and challenges of a maturing technology. Nat Rev Genet 10:669–680.

Petryszak R, Burdett T, Fiorelli B, Fonseca NA, Gonzalez-Porta M, Hastings E, Huber W, Jupp S, Keays M, Kryvych N, McMurry J, Marioni JC, et al. 2014. Expression Atlas update—a database of gene and transcript expression from microarray- and sequencing-based functional genomics experiments. Nucleic Acids Res 42:D926–D932.

Renkema KY, Stokman MF, Giles RH, Knoers NVAM. 2014. Next-generation sequencing for research and diagnostics in kidney disease. Nat Rev Nephrol 10:433–444.

Rossetti S, Hopp K, Sikkink RA, Sundsbak JL, Lee YK, Kubly V, Eckloff BW, Ward CJ, Winearls CG, Torres VE, Harris PC. 2012. Identification of Gene Mutations in Autosomal Dominant Polycystic Kidney Disease through Targeted Resequencing. J Am Soc Nephrol JASN 23:915–933.

Sarwal M, Sigdel T. 2013. A common blood gene assay predates clinical and histological rejection in kidney and heart allografts. Clin Transpl 241–247.

Sasagawa Y, Nikaido I, Hayashi T, Danno H, Uno KD, Imai T, Ueda HR. 2013. Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity. Genome Biol 14:R31.

Schumacher S, Muekusch S, Seitz H. 2015. Up-to-Date Applications of Microarrays and Their Way to Commercialization. Microarrays 4:196–213.

Shah SP, Roth A, Goya R, Oloumi A, Ha G, Zhao Y, Turashvili G, Ding J, Tse K, Haffari G, Bashashati A, Prentice LM, et al. 2012. The clonal and mutational evolution spectrum of primary triple negative breast cancers. Nature 486:.

Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, Seifert M, Borodina T, Soldatov A, Parkhomchuk D, Schmidt D, O'Keeffe S, et al. 2008. A Global View of Gene Activity and Alternative Splicing by Deep Sequencing of the Human Transcriptome. Science 321:956–960.

Suthanthiran M, Schwartz JE, Ding R, Abecassis M, Dadhania D, Samstein B, Knechtle SJ, Friedewald J, Becker YT, Sharma VK, Williams NM, Chang CS, et al. 2013. Urinary-cell mRNA profile and acute cellular rejection in kidney allografts. N Engl J Med 369:20–31.

Tan AY, Michaeel A, Liu G, Elemento O, Blumenfeld J, Donahue S, Parker T, Levine D, Rennert H. 2014. Molecular diagnosis of autosomal dominant polycystic kidney disease using next-generation sequencing. J Mol Diagn 16:216–228.

Taussig DC, Vargaftig J, Miraki-Moud F, Griessinger E, Sharrock K, Luke T, Lillington D, Oakervee H, Cavenagh J, Agrawal SG, Lister TA, Gribben JG, et al. 2010. Leukemia-

initiating cells from some acute myeloid leukemia patients with mutated nucleophosmin reside in the CD34- fraction. Blood 115:1976–1984.

Todd R, Kuo MWLWP. 2002. Gene expression profiling using laser capture microdissection. Expert Rev Mol Diagn 2:497–507.

Trujillano D, Bullich G, Ossowski S, Ballarín J, Torra R, Estivill X, Ars E. 2014. Diagnosis of autosomal dominant polycystic kidney disease using efficient PKD1 and PKD2 targeted next-generation sequencing. Mol Genet Genomic Med 2:412–421.

Wang Z, Gerstein M, Snyder M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10:57–63.

Yang L, Besschetnova TY, Brooks CR, Shah JV, Bonventre JV. 2010. Epithelial cell cycle arrest in G2/M mediates kidney fibrosis after injury. Nat Med 16:535–143.

Abstract

Renal diseases have a high impact on the economy of any health care system worldwide. In addition, patient numbers are steadily increasing over the past decades with a prevalence of over 500.000 new end stage renal disease (ESRD) worldwide cases every year. ESRD is the final stage of chronic kidney disease (CKD) that has as the leading causes diabetes and hypertension, as well as glomerulonephritis, urolithiasis, autosomal dominant polycystic kidney disease (ADPKD), and progression of acute kidney injury (AKI), among others. However, in many cases, the mechanisms of these diseases affecting kidney and its function are poorly understood, or difficult to diagnose. Within this study, we used newer technologies, methodologies, and data analysis approaches to throw some light into the pathomechanisms of CKD and AKI. Moreover, potentially improving the diagnostic value for already existing diagnostic assays (e.g. ADPKD).

In the past years, advances in DNA sequencing technologies have revolutionized the field of clinical research and diagnostics. High throughput sequencing such as next-generation sequencing (NGs) is being used because of its high quality and accuracy when analysing DNA samples. Other sequencing technologies have also shown their value such as long-read sequencing which is used because of its longer sequencing reads and accuracy resolving low-complexity sequences, such as repetitive regions or high GC-percent regions. Within the scope of this thesis we used several cutting-edge sequencing approaches applied to renal disease's clinical research to:

1. Improve the diagnostic value of already existing diagnostic assays for ADPKD. ADPKD is an inherited disease that accounts for 5% to 10% of ESRD. However, the screening of the main ADPKD gene *PKD1* is challenging because of its multi-exon structure, allelic heterogeneity, and high homology with six *PKD1* pseudogenes, as well as extremely high GC content. Using direct long-read sequencing we showed that ADPKD diagnostics without interference of *PKD1* homologous sequences is possible.
2. Characterize the expression profile of AKI and the underlying mechanisms using RNA sequencing. Patients undergoing major surgery may develop AKI which has been associated with higher mortality risk and reduced renal functionality, and high risk of progression of CKD. Some evidences pointed out to the tubular system being at the middle of this pathophysiology and further recovery. However, the factors

involved in this recovery are still poorly understood. In this context, we characterized renal messenger RNA expression profiles of AKI in wild type and Gdf15 knock out constructs. Gdf15 was identified to be associated with lower tubular damage suggesting a protective role of the kidney. We identified 89 transcription factors that are potentially driving the response mechanisms in AKI, as well as other 13 transcription factors possibly linked with the protective mechanisms of Gdf15.

3. Evaluate the practical boundaries of RNA sequencing to characterize glomerular diseases for CKD. The analysis of renal biopsies is very informative to determine patient's glomerular disease stage and progression rates. However, fresh frozen biopsies are limited or non-existent compared to the more abundant formalin-fixed, paraffin-embedded (FFPE) renal biopsies. FFPE tissue can be easily stored for long periods of time, allowing large sample archives with many years of clinical data collection and follow-up. We showed that characterizing glomerular disease expression profiles by RNA sequencing from FFPE samples is possible. However, our data suggests that the required number of glomeruli in cross sections may be higher than the number glomeruli present in a usual renal biopsy.

Finally, we elaborated about the future impact of the results obtained in the context of clinical research, and their value for the understanding or diagnostics of renal diseases.

Résumé

Les maladies rénales ont un impact important sur l'économie de tout système de santé dans le monde. En outre, le nombre de patients augmente régulièrement au cours des dernières décennies avec une prévalence de plus de 500 000 nouveaux cas de maladie rénale en phase terminale (ESRD) dans le monde entier chaque année. L'ESRD est l'étape finale de la maladie rénale chronique (CKD) qui a comme principales causes le diabète et l'hypertension, ainsi que la glomérulonéphrite, urolithiasis, la polykystose rénale autosomique dominante (ADPKD) et la progression de la lésion rénale aiguë (LRA), entre autres. Cependant, dans de nombreux cas, les mécanismes de ces maladies affectant le rein et sa fonction sont mal connus ou difficiles à diagnostiquer. Dans le cadre de cette étude, nous avons utilisé des technologies plus récentes, des méthodologies et des approches d'analyse de données pour jeter un peu de lumière dans les pathomécanismes de la CKD et de l'AKI. En outre, l'amélioration potentielle de la valeur diagnostique des tests diagnostiques déjà existants (par exemple ADPKD).

Au cours des dernières années, les progrès dans les technologies de séquençage de l'ADN ont révolutionné le domaine de la recherche clinique et du diagnostic. Le séquençage à haut débit tel que le séquençage de prochaine génération (NG) est utilisé en raison de sa haute qualité et de précision lors que l'analyse des échantillons d'ADN. D'autres technologies de séquençage ont également montré leur valeur, comme le séquençage à longue lecture qui est utilisé en raison de ses longues lectures de séquençage et de la précision de résolution de séquençages de faible complexité, telles que les régions répétitives ou des régions de GC-pourcentage élevé. Dans le cadre de cette thèse, nous avons utilisé plusieurs méthodes de pointe de séquençage appliquées à la recherche clinique sur la maladie rénale afin de:

1. Améliorer la valeur diagnostique des tests diagnostiques déjà existants pour l'ADPKD. ADPKD est une maladie héréditaire qui représente de 5% à 10% de l'ESRD. Cependant, le criblage du principal gène ADPKD PKD1 est difficile en raison de sa structure multi-exon, de son hétérogénéité allélique et de son homologie élevée avec six pseudogènes PKD1, ainsi que d'une teneur en GC extrêmement élevée. En utilisant le séquençage direct à longue lecture, nous avons montré que le diagnostic ADPKD sans interférence des séquences homologues PKD1 est possible.
2. Caractériser le profil d'expression de l'IRA et des mécanismes sous-jacents en utilisant le séquençage de l'ARN. Les patients qui subissent une chirurgie majeure peuvent

développer une IRA qui a été associée à un risque de mortalité plus élevé et une fonctionnalité rénale réduite, et un risque élevé de progression de la CKD. Certaines preuves indiquent que le système tubulaire est au milieu de cette pathophysiologie et de la récupération ultérieure. Cependant, les facteurs impliqués dans cette reprise sont encore mal compris. Dans ce contexte, nous avons caractérisé les profils d'expression de l'ARN messager rénal d'AKI dans des constructions de type sauvage et knock-out Gdf15. Gdf15 a été identifié comme étant associé à des lésions tubulaires inférieures suggérant un rôle protecteur du rein. Nous avons identifié 89 facteurs de transcription qui sont potentiellement moteurs des mécanismes de réponse dans l'IRA, ainsi que d'autres facteurs de transcription 13 éventuellement liés avec les mécanismes de protection de Gdf15.

3. Évaluer les limites pratiques du séquençage de l'ARN pour caractériser les maladies glomérulaires pour la CKD. L'analyse des biopsies rénales est très informative pour déterminer le stade de la maladie glomérulaire du patient et les taux de progression. Cependant, les biopsies congelées fraîches sont limitées ou inexistantes par rapport aux biopsies rénales fixées au formol et à la paraffine (FFPE) plus abondantes. Les tissus FFPE peuvent être facilement stockés pendant de longues périodes, ce qui permet de disposer d'importantes archives d'échantillons avec de nombreuses années de collecte de données cliniques et de suivi. Nous avons montré que la caractérisation des profils d'expression de la maladie glomérulaire par séquençage d'ARN à partir d'échantillons de FFPE est possible. Cependant, nos données suggèrent que le nombre requis de glomérules dans les coupes transversales peut être supérieur au nombre de glomérules présents dans une biopsie rénale habituelle.

Enfin, nous avons développé l'impact futur des résultats obtenus dans le cadre de la recherche clinique et leur valeur pour la compréhension ou le diagnostic des maladies rénales.

Acknowledgements

First, I would like to thank everyone that participated in any way in the development of this thesis and helped making this work possible.

I thank my supervisors, Joost and Bart, for their continued encouragement and support. Their guidance and trust taught me different ways of understanding science, which also included flexibility and creativity.

Besides the supervisors I would like to thank the Thesis Committee, Alberto and Zelmina, for their dedication and technical support.

My sincere thanks to all collaborators that made the project possible. Personally, I thank Dorien, Alberto, Lola, Yahya and Monique for their contribution in the form of samples, experience, and guidance.

I thank my fellow consortium and company colleagues, since they transformed all meetings into enriching nice and lively discussions. Particularly, Theo for his yearly assistance with the registration; Kirsten and Zoraide for the scientific, and not scientific, discussions; André for coffee talks about rose gardens.

Last, I would like to thank my friends and family for their unconditional support, and personally to Mireia for the encouragement, and valuable scientific and personal advice; to Carme for showing immense patience on difficult moments.