Universiteit
Leiden
The Netherlands

Data-driven knowledge discovery in polycystic kidney disease

Malas, T.

Citation
Malas, T. (2021, March 30). *Data-driven knowledge discovery in polycystic kidney disease*. Retrieved from https://hdl.handle.net/1887/3158169

Version:	Publisher's Version
License:	Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden
Downloaded from:	https://hdl.handle.net/1887/3158169

Note: To cite this publication please use the final published version (if applicable).

Cover Page

Universiteit Leiden

The handle https://hdl.handle.net/1887/3158169 holds various files of this Leiden University dissertation.

Author: Malas, T.
Title: Data-driven knowledge discovery in polycystic kidney disease
Issue Date: 2021-03-30

DATA-DRIVEN KNOWLEDGE DISCOVERY IN POLYCYSTIC KIDNEY DISEASE

Tareq Malas

Data-driven knowledge discovery in polycystic kidney disease

Tareq Malas
Leiden University Medical Center, The Netherlands

ISBN: 978-94-6423-160-1
Layout & design: Tareq Malas

© 2021, Tareq Malas. Copyright of the published material in chapters 2-5 lies with the publisher of the journal listed at the beginning of each chapter. All rights reserved. No part of this thesis may be reprinted, reproduced or utilized in any form by electronic, mechanical, or other means now known or hereafter invented, including photocopying and recording in any information storage or retrieval system without prior written permission of the author

DATA-DRIVEN KNOWLEDGE DISCOVERY IN POLYCYSTIC KIDNEY DISEASE

Proefschrift

ter verkrijging van
de graad van doctor aan de Universiteit Leiden,
op gezag van rector magnificus prof.dr.ir. H. Bijl,
volgens besluit van het college voor promoties
te verdedigen op dinsdag 30 maart 2021
klokke 11:15 uur

door

Tareq Malas

geboren in Damascus, Syrië

in 1987

Promotores:

Prof. Dr. D.J.M. Peters

Prof. Dr. P.A.C. ‘t Hoen

Promotiecommissie:
Prof. dr. ir. J.A.P. Willems van Dijk
Prof. dr. I. Meulenbelt
Prof. dr.ir. C.T.A. Evelo[1]
Dr. F. van Eeden[2]

The studies described in this thesis have been performed at the Department of Human Genetics, Leiden University Medical Center, the Netherlands.

1. Faculty of Health, Medicine and Life Sciences, Maastricht University

2. Department of Biomedical Science, University of Sheffield

Table of Contents

CHAPTER 1
GENERAL INTRODUCTION

1. Polycystic Kidney Disease, a common disorder needing effective treatment

A. Clinical Aspects of the Disease

Polycystic Kidney Disease (PKD) is a genetic disease of the kidney with 4/10000 prevalence[1]. PKD is characterized by the gradual replacement of normal kidney parenchyma by fluid-filled cysts and fibrotic tissue. Renal cysts grow and increase with age, leading to distortion of normal kidney architecture and ultimately, end-stage renal disease (ESRD) in a majority of patients. Clinical manifestations include a higher frequency of intracranial aneurysms (ICAs) and severe polycystic liver disease (PCLD), for which resection or other surgery may be required[2]. Other symptoms that PKD patients suffer from include, pain or tenderness in the abdomen, frequent urination, fatigue, kidney stones, and others. Autosomal Dominant Polycystic Kidney Disease (ADPKD) is caused by mutations in the *PKD1* or *PKD2* gene and the less frequent autosomal recessive form, Autosomal Recessive Polycystic Kidney Disease (ARPKD), is caused by mutations in the *PKHD1* gene[1,3]. PKD has a varying age of onset, where 50% of patients develop end-stage kidney disease by the age of 50[4].

Diagnosis of ADPKD is mainly performed by renal imaging such as ultrasonography, computed tomography, or magnetic nuclear resonance[5] and molecular diagnostics are necessary when a definite diagnosis is required. PKD's molecular diagnostics are challenging, especially for ADPKD where the *PKD1* gene is difficult to sequence. Approximately 70% of the 5′ genomic region of the *PKD1* gene is duplicated six times on chromosome 16p within six pseudogenes (*PKD1P1 to PKD1P6*), which share a 97.7% sequence identity with the genuine gene[6,7]. The high GC content at the gene locus can also bias sequencing since more (or less) sequencing reads tend to come from a region with a higher GC content. Additional complications in sequencing the *PKD1* gene arise from the presence of many missense variants, the absence of mutation hotspots, and the high allelic heterogeneity of ADPKD. As personalized medicine gains popularity, the demand for simple and cost-effective molecular approaches will increase.

B. The role of Polycystins and the two-hit model in cyst formation

PKD1 and *PKD2* protein products, Polycystin-1 (PC1) and Polycystin-2 (PC2). have been extensively studied, however their exact role in cyst formation is not yet fully understood. PC1 regulates signaling pathways essential for proper tubular structures in kidney and liver[8-12] and suggests that a threshold level might be required to prevent cyst formation[13,14]. Cystogenesis will begin when the level of functional PC1 is below the critical threshold[15,16], and the degree at which the PC1 activity levels drop below this threshold determines speed of cyst formation and ADPKD severity. A fundamental property of PC1 is its post-translational modification by cleavage at the juxtamembrane GPS motif[17], defective cleavage is thought to play a significant role in ADPKD's pathogenesis. Additionally, PC1 C-terminal tail (PC1-CTT) has important signaling implications. Recently It was shown to regulate the complement factor B expression by signal transducer and activator of transcription 1[18]. The PC1-CTT was also shown to associate with β-catenin and act as an inhibitor of Wnt-dependent intracellular signaling, a signaling pathway that promotes epithelial cell

proliferation and found upregulated in PKD patients[19]. However, the exact cellular functions of PC1 is yet to be fully understood. PC2 is a TRP-nonselective, Ca2+-permeable cation channel[20] regulated by diverse stimuli including divalent cations, pH, voltage and phosphorylation[20]. PC1 has been hypothesized to form a mechanosensitive cation channel complex with PC2 in the primary cilia[21-23]. Functional defects in this complex caused by mutation of *PKD1* or *PKD2* result in autosomal dominant polycystic kidney disease (ADPKD)[24,25]. The cilium is a microtubule-based organelle found on most cells in the mammalian body. In the kidney, the primary cilia is present on most cells of the nephron and extend off the apical surface of the epithelium into the tubule lumen. It is now thought that the primary cilia senses fluid flow through the lumen of renal tubules by acting as a mechanosensor and initiating a cascade of downstream molecular signaling events. Altered signaling as a result of defective cilia function due to *PKD1/PKD2* mutations is hypothesized to trigger cyst formation[22,26].

Despite this progress in understanding the functions of polycystins, the primary cause of cyst formation remains elusive. Understanding the main cause of cyst formation will enable targeting the primary rather than downstream secondary mechanisms which is likely to be more effective. Currently, the most widely accepted theory for cyst generation in human PKD is the “two-hit hypothesis.” ADPKD patients are typically heterozygous, with one PKD allele having a germline mutation (first hit) and the other is normal. The remaining normal *PKD1/PKD2* allele develops a somatic mutation (second hit) in a small percentage of the cells. The “two-hit hypothesis” was demonstrated when the epithelial cells lining a human cyst were isolated and confirmed to be monoclonal, and found to have a loss of heterozygosity at the *PKD1* locus (normal haplotype was lost)[27].

C. Targeting the Signaling Pathways Involved in PKD

There are many signaling pathways that appear to be compromised in PKD. These include activator protein-1 (AP-1) transcription factor, G-coupled protein receptors (GPCR), B-Raf/ERK, mitogen-activated protein kinase (MAPK), EGFR signaling, mammalian target of rapamycin (mTOR), as well as second messengers like cAMP and Ca2+ (Figure-1). PC1 directly binds the G-protein α-subunits and lead to the activation of subsequent signaling pathways such as AP-1 transcription factor, c-Jun N-kinase, and the nuclear factor of activated T-cell signaling cascade, which in turn regulate cell proliferation, differentiation and apoptosis[28] (Figure-1). mTOR is a serine/threonine protein kinase that is involved in the regulation of cell proliferation, cell metabolism, protein synthesis and transcription. The polycystins, in part regulate the mTOR signaling pathway, as mTOR was shown to be increased in PKD[29]. In normal condition, PC1 inhibits mTOR signaling by stabilizing the TSC1/TSC2 complex, which is required for mTOR to function[30]. Interestingly, inhibitors of mTOR were shown to slow cyst formation at least in preclinical models of PKD[31,32]. Two large randomized clinical trials testing the mTOR inhibitors sirolimus and everolimus in ADPKD patients failed to slow the progression of the disease[33,34]. Cyclic AMP (cAMP), a second messenger involved in various cellular processes, including cell proliferation and differentiation, is elevated in human and animal model PKD[35-37]. By stimulating epithelial cell proliferation, cAMP is known to promote cyst development[38,39]. Strategies that focus on lowering cAMP levels have been successful in slowing cyst formation in animal and human PKD models. Somatostatin that works by inhibiting cAMP accumulation was shown to be effective in slowing progression in liver and kidney cystic disease in a rat model of PKD[40].

The somatostatin analogue octreotide was also shown to be effective in reducing kidney volume in ADPKD patients[41,42], and pravastatin is currently undergoing clinical trials for its effect on slowing cyst formation in young adults with ADPKD[43]. Additionally, the vasopressin V2 receptor antagonists tolvaptan that reduced renal cyclic AMP levels, inhibited renal cystogenesis and kidney enlargement[44] and now is approved for therapy in Europe, Canada and Japan. Epidermal growth factor (EGF) plays an important role in cyst epithelial cell proliferation and cyst expansion. Inhibition of the epidermal growth factor receptor was successful in reducing cyst formation in a number of animal models of PKD[45,46]. Receptor tyrosine kinase inhibitors are also showing success in slowing the progression of PKD, in particular tesevatinib, is currently undergoing phase-2 clinical trials for ADPKD patients and had positive results in rodent models of autosomal recessive polycystic kidney disease[47]. Prominent defective metabolic rates have also been described in ADPKD animal models, providing additional opportunities for therapy. Modulation of the metabolic processes in PKD models either via diet-restriction or inhibition of glycolysis resulted in ameliorating the kidney volume, cystic index and reduced proliferation rates[48-50].

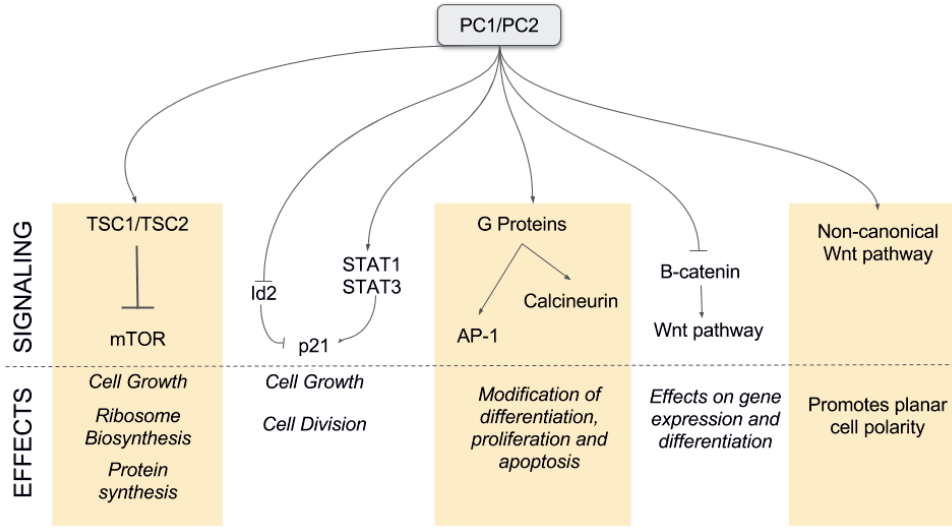

Figure-1: Overview of the signaling pathways affected by PC1 and PC2. PC1 and PC2 affect multiple signaling pathways. Figure-1 presents a summary of the signaling pathways influenced by PC1 and PC2. These pathways directly or indirect stimulate or inhibit different aspects of cellular growth and differentiation[51].

D. Renal Injury aggravates PKD

Several studies indicated a tissue injury repair component in the pathology of PKD[52-54]. Indeed, there are numerous similarities between PKD and renal injury, since both phenotypes are accompanied by a combination of processes including proliferation, secretion of growth factors as well as inflammation. Weimbs proposed a model where *Polycystin-1* (*PC1)*, the protein encoded by *PKD1*, and primary cilia have a critical function in sensing renal injury, by detecting changes in luminal fluid flow, and triggering proliferation[29]. Besides a proposed mechanistic overlap, several studies showed that renal injury could stimulate cyst progression. For example, nephrotoxic injury in an ADPKD adult mouse model resulted in accelerated cyst formation and a more progressive phenotype[52]. This is further supported by findings that ischemic reperfusion injury and also tubular cell hypertrophy following unilateral nephrectomy accelerated PKD[52,55,56]. Although the link between PKD and renal injury seems rather strong, until now a thorough comparison between the two conditions at the molecular level has not been made, and little is known about the key genes and pathways shared between the two.

2. Transcriptomics revolutionized the way we study diseases

A. Introduction of Transcriptomic Technologies

A transcriptome is the full range of RNA molecules expressed by an organism or a cell. This word appeared first in an article by Velculescu *et al* when they published the yeast transcriptome using Serial analysis of gene expression (SAGE) in 1997[57]. The SAGE technology allowed for the cloning of short cDNA transcript fragments that are subsequently sequenced by Sanger sequencing and counted[58]. SAGE provided a comprehensive expression profiling for a given cell population unleashing the age of high-throughput expression profiling. The invention of the microarray technology in the mid 2000s allowed the assay of thousands of transcripts simultaneously at a greatly reduced cost per gene and labour savings[59]. Microarrays are a collection of microscopic DNA spots attached to a solid surface. The principle behind microarrays relies on the hybridization between two DNA strands. Complementary nucleic acid sequences pair forming hydrogen bonds. Microarrays use fluorescently labeled target sequences, which pass over the microarray probes and bind to complementary sequences generating a signal. The strength of this signal depends on the amount of target sample binding to the probes. Using relative quantitation, microarrays determine which probes are significant. The rapid development of high-throughput sequencing technologies gave rise to RNA-Seq. RNA-Seq refers to high-volume sequencing of cDNA transcripts. The key advantage of RNA-Seq over hybridization-based microarrays is the depth and novelty of the output based on unbiased sequence information[60]. The first application of RNA-Seq was published in 2006 with 10^5 transcripts sequenced[61], now sequencing depth can exceed 10^9 [62,63]. This vast increase in yield makes it possible to accurately quantify the entire human transcriptome. It is now appreciated that 85% of the human genome can be transcribed and only 3% of it encodes protein-coding genes[64]. Thus, RNA-Seq has been instrumental to study the diversity of novel transcript species including long noncoding RNA, mircoRNAs, siRNA, and other small RNA classes (eg, snRNA and piRNA). These RNA species are gaining more importance in disease studies because of their involvement in regulating RNA stability, protein translation and modulation of chromatin states[65]. For instance, RNA-Seq has been used to discover enhancer RNAs, a class of short

transcript directly transcribed from the enhancer region, which contributes to our knowledge of epigenetic gene regulation[66,67]. In addition, RNA-Seq can give information about transcriptional start sites, revealing alternative promoter usage, information about alternative splicing, and transcription termination at the 3′ end, which is critical for mRNA stability[68,69]. RNA-Seq data essentially do not display background signal like microarrays, because transcripts that are not there will not be sequenced. However, background signal may arise, *e.g.* in pseudogenes, because ambiguous mapping of sequence reads in the genome (multi-mapping). Since RNA-Seq is based on the number of sequences mapped there is no upper limit to its quantification. In contrast, microarrays lack sensitivity for genes expressed at very high or low levels. Continuous developments in the RNA-Seq technology allowed for pair-end sequencing, strand-specific sequencing and single-cell sequencing.

Several limitations remain in the field of transcriptomics. The ideal method for transcriptomics should be able to directly identify and quantify all RNAs, small or large. However cDNA library construction is still required in RNA-Seq. This includes RNA fragmentation, reverse transcription, and PCR amplification. Each of these steps comes with biases that skew the representation of different RNA species in the sequencing data[70]. Another important aspect in transcriptomics is sequence coverage which correlates with increased costs. To detect rare transcripts and all possible isoforms in human samples, considerable sequencing depth (*i.e.* >200 million paired-end reads) is required[71]. Additionally, data processing, storage, management and analysis are still a major bottleneck in transcriptomics.

The transcriptomics field is expected to continue to develop and progress. The next big challenge for transcriptomics lies in data interpretation and the production of actionable insights in the upfront of medical and life sciences applications. Transcriptomics is not the only evolving *omics* technology. We are witnessing improvements in epigenetic profiling, proteomics and metabolomics as well. Each of the *omics* contributes a piece of information that is crucial for biomedical research. Proteomics complements transcriptomics by quantifying all proteins that are expressed, and modified following expression in a cell or a given tissue. Relying only on transcriptomics will miss the wide variety of chemical modifications *i.e.* phosphorylation and ubiquitination proteins undergo after translation. Many of these post-translational modifications are critical to the protein's function. Integrating transcriptomics with other *omics* is critical to our understanding of diseases and enabling the process of drug discovery.

B. Applications of Transcriptomics

Transcriptomics has a wide range of applications across diverse areas of biomedical research. In disease diagnosis and profiling, RNA-Seq has allowed for the identification of transcriptional start sites at a large-scale and revealed novel alternative splicing events. Defining these variants is critical to the interpretation of disease association studies[72] . Additionally, RNA-Seq is being used to study (allele-specific) gene expression regulation, identify disease-associated single nucleotide variants(SNVs) and somatic mutations, gene fusions, and RNA editing[73-75]. RNA-Seq is leading the way in expression profiling studies, where comparison of different disease states or of disease and control samples are becoming a commonplace in biomedical research. This has given rise to a wealth of information on molecular pathways and gene co-expression networks. For example the

application of gene co-expression network analysis, that identifies modules of genes expressed in a similar pattern, was applied to study the genes involved in several diseases, such as autism and cancer[76,77]. The identification of such modules and/or molecular pathways are changing drug discovery methods. Scientists are interested Identifying drugs that target key proteins of key expression modules and molecular pathways to discover novel drugs for a number of diseases[78,79].

C. Transcriptomics of Polycystic Kidney Disease

In renal diseases, transcriptomics was applied to study acute renal injury, chronic renal disease and polycystic kidney disease. Microarrays were the main platform used in the study of renal diseases. In these attempts, samples from patients and disease models were compared to wildtype controls and genes that were expressed differently were identified and further analyzed. In principle, the differentially expressed genes (DEGs) resulting from expression profiling experiments would reflect the genes involved in a certain disease and condition. However, DEGs can also arise as a result of differences in tissue, cellular composition and other experimental biases not related to the used expression capturing platform. In PKD, expression profiling experiments were attempted by several groups[80-88]. The depth and scope of these studies varied considerably. The biggest difference between the studies was the sample of origin used in the analysis. Studies varied from using cell lines, patient-derived material, and whole kidneys from rats or mice. Each type of sample provides a unique set of advantages and challenges. Human patient samples would seem ideal since they best reflect the disease under investigation. However, the disease is characterized by cysts arising from every nephron segment, so comparing “cystic” vs. “control” patient material could mean comparing different nephron segments, a fact rarely acknowledged or explicitly controlled for. Additionally, extracting the expression profiles from advanced stages of the disease may superimpose additional changes not related to the cause of cyst formation. Such changes can be the result of uremia and renal injury. Additional differences between the studies include sample size, technology used and downstream analysis of the resulting data. All of these factors have a large influence on the results obtained from the expression profile experiments. As one would expect from the many experimental differences between the different expression profiles performed on polycystic kidney disease, different conclusions were reached and a long list of pathways was suggested to be disease-related. Such pathways include extracellular matrix defects[83], epithelial-to-myofibroblast transition[85], apolipoprotein expression[86], RXR pathways[81], and various miscellaneous, broad functional categories involving signaling, metabolic and developmental pathways[82,89,90].

3. Data explosion and the need for proper integration

A. Historical and Projected Trends in Data Growth

With the advent of high-throughput omics technologies, life scientists are continuously generating large volumes of data (Figure-2). New technologies such as RNA-Seq, are making it easier and cheaper to perform experiments that generate large quantities of data (Figure-2). Genomics data are currently being produced at an unprecedented rate, doubling every seven months[91]. In fact, it is expected that data resulting from sequencing technologies only will reach more than exabase (1000^6) of sequence per year in the next five

years and approach one zettabase (1000^7) of sequence per year by 2025[91]. This exponential growth in sequencing data is fueled by personalized medicine, large population sequencing projects, single cell genome sequencing projects and others. Additional sources of large data in biomedical research will come from other *omic*-technologies like proteomics and imaging.

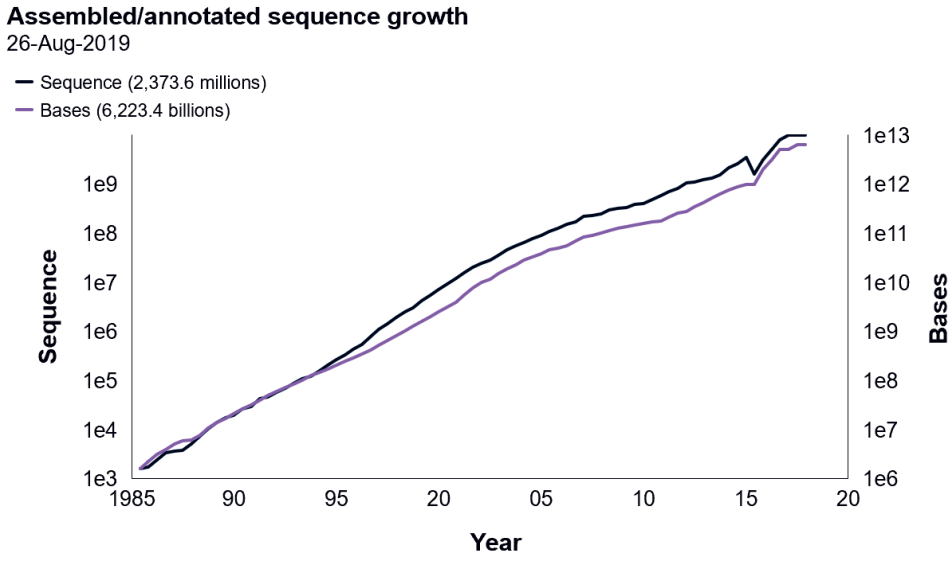

Figure-2: Number of sequences and bases deposited in the European Nucleotide Archive. The amount of sequences and DNA bases deposited in the European Nucleotide Archive – adopted from The European Nucleotide Archive in 2019.

B. Challenges in Data Explosion

Life sciences is one of many other fields that have to deal with the emergence of big data. Other fields include, astronomics, education, banking and social media. According to IBM, big data are characterized by their volume, variety, velocity and veracity, also known as the four V's of Big Data. The volume of data refers to the scale of big data. In life sciences, data is expected to grow exponentially reaching the magnitude of zettabytes in the near future. This dictates that storage capacity needs to be increased drastically to meet this vast increase in data. On the other hand improvements in data compression techniques and storage methods will become critical for data management. The properties of DNA sequence offer a chance to build DNA specific compression algorithms. One such technique is a reference-based compression method that efficiently compresses DNA sequences by comparing the genomic sequence to a reference genome and looking for differences[92]. Variety refers to the different forms of generated data, In life sciences data can come in many formats. In additions to *omics*, data in life sciences come in the form of text and tables in medical records and publications, multi-media from medical appliances and research related experiments and now data from social media of patient/doctor communities and personal medical devices. Each format of the aforementioned has its own challenges in terms of storage, processing and sharing. Data interoperability that relates to the ability of

systems and services to have clear expectations for the contents, context and meaning of the exchanged data, is a major challenge for the variety of data produced and consumed in life sciences. The velocity (speed of data generation) is at an ever increasing pace, now real-time generated patient/doctor data is becoming a norm and immediate processing a must. The veracity refers to the certainty and the reliability of the data. As datasets continue to grow in size in life sciences, the presence of noisy data increases. This issue is of particular importance in medicine in which evidence-driven decisions are the foundation of patient care.

To manage the requirements for the processing and storage of big data, solutions now apply the divide and conquer strategy. The idea is to partition a large problem into more tractable and independent subproblems[93]. Each subproblem is tackled in parallel by different processing units. In small scale, such divide-and-conquer paradigm can be implemented either by multi-core computing or grid computing. However the scalability of this solution is limited for the basic assumption, that at least one of the many different nodes is deemed to malfunction at one point. To solve this problem, big data algorithms copy the same data chunk to more than one node, making it available in case the other node failed [building in redundancy]. Cloud computing, defined as the practice of using a network of remote servers hosted on the Internet to store, manage, and process data, rather than a local server or a personal compute [Oxford Dictionaries:2017], are emerging as a feasible solutions for life sciences. Large cloud computing service providers such as Amazon are providing readily available infrastructure for the biotech and pharma industries. Data transfer speed remains a limiting factor despite the improvements in network infrastructure and more so in the less developed countries. According to Akamai Q4 2015m the global average internet speed is 5.6 Mb/s at which a 1GB file will take approximately 26 minutes to download. Other limitations to cloud computing include privacy concerns and copyright issues.

C. Policies in Data Management

Data management is a critical aspect of maintaining a reliable data source, this is true for both large and small datasets. However, data management is gaining increased attention in the big data era. This is because incomplete data management policies that would work for small datasets are not feasible/applicable for large datasets. Data are considered as a valuable resource for industries and universities alike, and data loss or data silos are no longer accepted. There is an increasing need for data sharing to exploit most of the potential information lying within large datasets for added value and knowledge discovery. These rapid changes in how data is perceived are driving all data stakeholders to establish regulations and policies that direct how data should be stored, processed and shared.

One such effort is seen in the The FAIR Data Principles, which are a set of guidelines put together by a team of stakeholders, representing academia, industry, funding agencies, and scholarly publishers, aimed at those wishing to enhance the reusability of their data holdings[94]. The four foundational principles of FAIR are, Findability, Accessibility, Interoperability, and Reusability. The FAIR Principles emphasize the ability of machines to automatically find and use the data. Importantly, these principles must apply not only to ‘data’ in the conventional sense, but also to the algorithms, tools, and workflows that led to that data[94]. In addition, data must be accompanied by a detailed and descriptive metadata section.The metadata is crucially needed to describe the data and how it was generated, as

well as putting it in context of other data sources. Data sources that adhere to the FAIR principles are easier to access, integrate and compare. Whilst, completely following the FAIR principles is challenging and can only be achieved gradually, databases can begin by adding metadata to their databases, modeling them into data structures that allow integration and querying. The Resource Description Framework (RDF), originally developed as a model to represent information about World Wide Web resources, is now commonly applied for the modeling of metadata and, to a lesser extent the data itself, in biomedical databases such as the *EMBL-EBI Expression Atlas*. RDF data can be queried using the SPARQL language [Prud, Eric, and Andy Seaborne. "SPARQL query language for RDF." (2006)]. Thus, the adoption of RDF across different biomedical databases is expected to facilitate the integration of life sciences data and allow scientists to devote more time on knowledge discovery. We are also witnessing the emergence of knowledge graphs in life sciences to integrate a large number of scattered databases. Knowledge graphs are networks of defined semantic concepts connected by edges based on a variety of resources. Edges can also be defined by semantically defined predicate types, that describe the type of the relationship connecting the two concepts. These graphs are read in the form subject-predicate-object and referred to as a semantic triple[95]. Examples of these efforts are the work of *Hettne et al* where they created a semantic knowledge graph of known drug disease associations, which was used to inference novel drug disease associations[96] and the work of *Chen et al* where they built a semantic knowledge graph for drug targets associations by integrating data from public datasets relating to drugs, chemical compounds, protein targets, diseases, side effects and pathways[97].

The growth in scientific data is expected to continue growing at an exponential rate. Innovations in data storage, processing and sharing are critical to unleash the full potential of the generated data for scientific discovery. Additionally, there is a pressing need for scientific communities to get together and establish principles for data management. We are seeing an increased reliance on big data in life sciences and the study of human genetics combined with gradual shift in policy mandating better storage and sharing of datasets. *Omic* technologies will continue to take the center stage in the study of diseases, and consolidated databases will emerge. In PKD, we are witnessing accelerated drug discovery methods that rely on big data and high-throughput experiments. As is the case in other industries, data will be the most valuable asset in life sciences research and development.

4. Aims and Outline of this Thesis

The aim of this thesis was to analyze Polycystic Kidney Disease (PKD) expression profiles to identify novel druggable gene targets and molecular pathways for its treatment.

Noise attributed to intrinsic differences across different disease models is known to dilute the main disease signal and complicate the study of the disease characteristics. In **chapter 2** we aimed at identifying a robust PKD Signature across different disease models and experimental setups. Using this signature, we aimed in **chapter 3** at identifying key transcriptional factors that impact cyst formation. Transcription factors are known to orchestrate the expression of a large number of genes making them interesting drug targets.

Since PKD is a progressive disease with varying molecular characteristics throughout its progression phases, in **chapter 4,** we aimed at identifying the key molecular pathways involved in each disease phase. We hypothesize that drugs targeting these key PKD genes

and molecular pathways will be important candidates for slowing down the cyst formation in PKD patients.

In **chapter 5** we evaluated drug discovery methods and in particular the possibilities of using semantic knowledge graphs for the discovery of novel drug disease associations and drug repurposing. Such graphs are interesting since they combine and integrate a large number of databases including literature text- mining of published abstracts.

REFERENCES

1 Audrezet, M. P. *et al.* Autosomal dominant polycystic kidney disease: comprehensive mutation analysis of PKD1 and PKD2 in 700 unrelated patients. *Hum Mutat* **33**, 1239-1250, doi:10.1002/humu.22103 (2012).

2 Pirson, Y., Chauveau, D. & Torres, V. Management of cerebral aneurysms in autosomal dominant polycystic kidney disease. *J Am Soc Nephrol* **13**, 269-276 (2002).

3 Bergmann, C. *et al.* PKHD1 mutations in autosomal recessive polycystic kidney disease (ARPKD). *Hum Mutat* **23**, 453-463, doi:10.1002/humu.20029 (2004).

4 Milutinovic, J. *et al.* Autosomal dominant polycystic kidney disease: symptoms and clinical findings. *Q J Med* **53**, 511-522 (1984).

5 Pei, Y. & Watnick, T. Diagnosis and screening of autosomal dominant polycystic kidney disease. *Adv Chronic Kidney Dis* **17**, 140-152, doi:10.1053/j.ackd.2009.12.001 (2010).

6 Rossetti, S. *et al.* Identification of gene mutations in autosomal dominant polycystic kidney disease through targeted resequencing. *J Am Soc Nephrol* **23**, 915-933, doi:10.1681/ASN.2011101032 (2012).

7 Bogdanova, N. *et al.* Homologues to the first gene for autosomal dominant polycystic kidney disease are pseudogenes. *Genomics* **74**, 333-341, doi:10.1006/geno.2001.6568 (2001).

8 Yu, S. *et al.* Essential role of cleavage of Polycystin-1 at G protein-coupled receptor proteolytic site for kidney tubular structure. *Proc Natl Acad Sci U S A* **104**, 18688-18693, doi:10.1073/pnas.0708217104 (2007).

9 Piontek, K., Menezes, L. F., Garcia-Gonzalez, M. A., Huso, D. L. & Germino, G. G. A critical developmental switch defines the kinetics of kidney cyst formation after loss of Pkd1. *Nat Med* **13**, 1490-1495, doi:10.1038/nm1675 (2007).

10 Piontek, K. B. *et al.* A functional floxed allele of Pkd1 that can be conditionally inactivated in vivo. *J Am Soc Nephrol* **15**, 3035-3043, doi:10.1097/01.ASN.0000144204.01352.86 (2004).

11 Boletta, A. & Germino, G. G. Role of polycystins in renal tubulogenesis. *Trends Cell Biol* **13**, 484-492, doi:10.1016/s0962-8924(03)00169-7 (2003).

12 Lu, W. *et al.* Perinatal lethality with kidney and pancreas defects in mice with a targetted Pkd1 mutation. *Nat Genet* **17**, 179-181, doi:10.1038/ng1097-179 (1997).

13 Jiang, S. T. *et al.* Defining a link with autosomal-dominant polycystic kidney disease in mice with congenitally low expression of Pkd1. *Am J Pathol* **168**, 205-220, doi:10.2353/ajpath.2006.050342 (2006).

14 Lantinga-van Leeuwen, I. S. *et al.* Lowering of Pkd1 expression is sufficient to cause polycystic kidney disease. *Hum Mol Genet* **13**, 3069-3077, doi:10.1093/hmg/ddh336 (2004).

15 Fedeles, S. V., Gallagher, A. R. & Somlo, S. Polycystin-1: a master regulator of intersecting cystic pathways. *Trends Mol Med* **20**, 251-260, doi:10.1016/j.molmed.2014.01.004 (2014).

16 Hopp, K. *et al.* Functional polycystin-1 dosage governs autosomal dominant polycystic kidney disease severity. *J Clin Invest* **122**, 4257-4273, doi:10.1172/JCI64313 (2012).

17 Qian, F. *et al.* Cleavage of polycystin-1 requires the receptor for egg jelly domain and is disrupted by human autosomal-dominant polycystic kidney disease 1-associated mutations. *Proc Natl Acad Sci U S A* **99**, 16981-16986, doi:10.1073/pnas.252484899 (2002).

18 Wu, M. *et al.* The C-terminal tail of polycystin-1 regulates complement factor B expression by signal transducer and activator of transcription 1. *Am J Physiol Renal Physiol* **310**, F1284-1294, doi:10.1152/ajprenal.00428.2015 (2016).

19 Lal, M. *et al.* Polycystin-1 C-terminal tail associates with beta-catenin and inhibits canonical Wnt signaling. *Hum Mol Genet* **17**, 3105-3117, doi:10.1093/hmg/ddn208 (2008).

20 Grieben, M. *et al.* Structure of the polycystic kidney disease TRP channel Polycystin-2 (PC2). *Nat Struct Mol Biol* **24**, 114-122, doi:10.1038/nsmb.3343 (2017).

21 Harris, P. C. & Torres, V. E. Polycystic kidney disease. *Annu Rev Med* **60**, 321-337, doi:10.1146/annurev.med.60.101707.125712 (2009).

22 Nauli, S. M. *et al.* Polycystins 1 and 2 mediate mechanosensation in the primary cilium of kidney cells. *Nat Genet* **33**, 129-137, doi:10.1038/ng1076 (2003).

23 Yoder, B. K., Hou, X. & Guay-Woodford, L. M. The polycystic kidney disease proteins, polycystin-1, polycystin-2, polaris, and cystin, are co-localized in renal cilia. *J Am Soc Nephrol* **13**, 2508-2516, doi:10.1097/01.asn.0000029587.47950.25 (2002).

24 Cornec-Le Gall, E., Audrezet, M. P., Le Meur, Y., Chen, J. M. & Ferec, C. Genetics and pathogenesis of autosomal dominant polycystic kidney disease: 20 years on. *Hum Mutat* **35**, 1393-1406, doi:10.1002/humu.22708 (2014).

25 Igarashi, P. & Somlo, S. Polycystic kidney disease. *J Am Soc Nephrol* **18**, 1371-1373, doi:10.1681/ASN.2007030299 (2007).

26 Praetorius, H. A. & Spring, K. R. Bending the MDCK cell primary cilium increases intracellular calcium. *J Membr Biol* **184**, 71-79, doi:10.1007/s00232-001-0075-4 (2001).

27 Qian, F., Watnick, T. J., Onuchic, L. F. & Germino, G. G. The molecular basis of focal cyst formation in human autosomal dominant polycystic kidney disease type I. *Cell* **87**, 979-987, doi:10.1016/s0092-8674(00)81793-6 (1996).

28 Parnell, S. C. *et al.* The polycystic kidney disease-1 protein, polycystin-1, binds and activates heterotrimeric G-proteins in vitro. *Biochem Biophys Res Commun* **251**, 625-631, doi:10.1006/bbrc.1998.9514 (1998).

29 Shillingford, J. M. *et al.* The mTOR pathway is regulated by polycystin-1, and its inhibition reverses renal cystogenesis in polycystic kidney disease. *Proc Natl Acad Sci U S A* **103**, 5466-5471, doi:10.1073/pnas.0509694103 (2006).

30 Distefano, G. *et al.* Polycystin-1 regulates extracellular signal-regulated kinase-dependent phosphorylation of tuberin to control cell size through mTOR and its downstream effectors S6K and 4EBP1. *Mol Cell Biol* **29**, 2359-2371, doi:10.1128/MCB.01259-08 (2009).

31 Shillingford, J. M., Piontek, K. B., Germino, G. G. & Weimbs, T. Rapamycin ameliorates PKD resulting from conditional inactivation of Pkd1. *J Am Soc Nephrol* **21**, 489-497, doi:10.1681/ASN.2009040421 (2010).

32 Tao, Y., Kim, J., Schrier, R. W. & Edelstein, C. L. Rapamycin markedly slows disease progression in a rat model of polycystic kidney disease. *J Am Soc Nephrol* **16**, 46-51, doi:10.1681/ASN.2004080660 (2005).

33 Braun, W. E., Schold, J. D., Stephany, B. R., Spirko, R. A. & Herts, B. R. Low-dose rapamycin (sirolimus) effects in autosomal dominant polycystic kidney disease: an open-label randomized controlled pilot study. *Clin J Am Soc Nephrol* **9**, 881-888, doi:10.2215/CJN.02650313 (2014).

34 Walz, G. *et al.* Everolimus in patients with autosomal dominant polycystic kidney disease. *N Engl J Med* **363**, 830-840, doi:10.1056/NEJMoa1003491 (2010).

35 Starremans, P. G. *et al.* A mouse model for polycystic kidney disease through a somatic in-frame deletion in the 5' end of Pkd1. *Kidney Int* **73**, 1394-1405, doi:10.1038/ki.2008.111 (2008).

36 Gattone, V. H., 2nd, Wang, X., Harris, P. C. & Torres, V. E. Inhibition of renal cystic disease development and progression by a vasopressin V2 receptor antagonist. *Nat Med* **9**, 1323-1326, doi:10.1038/nm935 (2003).

37 Hanaoka, K. & Guggino, W. B. cAMP regulates cell proliferation and cyst formation in autosomal polycystic kidney disease cells. *J Am Soc Nephrol* **11**, 1179-1187 (2000).

38 Parker, E. *et al.* Hyperproliferation of PKD1 cystic cells is induced by insulin-like growth factor-1 activation of the Ras/Raf signalling system. *Kidney Int* **72**, 157-165, doi:10.1038/sj.ki.5002229 (2007).

39 Hanaoka, K., Devuyst, O., Schwiebert, E. M., Wilson, P. D. & Guggino, W. B. A role for CFTR in human autosomal dominant polycystic kidney disease. *Am J Physiol* **270**, C389-399, doi:10.1152/ajpcell.1996.270.1.C389 (1996).

40 Caroli, A. *et al.* Reducing polycystic liver volume in ADPKD: effects of somatostatin analogue octreotide. *Clin J Am Soc Nephrol* **5**, 783-789, doi:10.2215/CJN.05380709 (2010).

41 Higashihara, E. *et al.* Safety study of somatostatin analogue octreotide for autosomal dominant polycystic kidney disease in Japan. *Clin Exp Nephrol* **19**, 746-752, doi:10.1007/s10157-014-1047-1 (2015).

42 Ruggenenti, P. *et al.* Safety and efficacy of long-acting somatostatin treatment in autosomal-dominant polycystic kidney disease. *Kidney Int* **68**, 206-216, doi:10.1111/j.1523-1755.2005.00395.x (2005).

43 Cadnapaphornchai, M. A. *et al.* Effect of pravastatin on total kidney volume, left ventricular mass index, and microalbuminuria in pediatric autosomal dominant polycystic kidney disease. *Clin J Am Soc Nephrol* **9**, 889-896, doi:10.2215/CJN.08350813 (2014).

44 Torres, V. E. *et al.* Tolvaptan in patients with autosomal dominant polycystic kidney disease. *N Engl J Med* **367**, 2407-2418, doi:10.1056/NEJMoa1205511 (2012).

45 Torres, V. E. *et al.* EGF receptor tyrosine kinase inhibition attenuates the development of PKD in Han:SPRD rats. *Kidney Int* **64**, 1573-1579, doi:10.1046/j.1523-1755.2003.00256.x (2003).

46 Sweeney, W. E., Jr. *et al.* Combination treatment of PKD utilizing dual inhibition of EGF-receptor activity and ligand bioavailability. *Kidney Int* **64**, 1310-1319, doi:10.1046/j.1523-1755.2003.00232.x (2003).

47 Sweeney, W. E., Frost, P. & Avner, E. D. Tesevatinib ameliorates progression of polycystic kidney disease in rodent models of autosomal recessive polycystic kidney disease. *World J Nephrol* **6**, 188-200, doi:10.5527/wjn.v6.i4.188 (2017).

48 Rowe, I. *et al.* Defective glucose metabolism in polycystic kidney disease identifies a new therapeutic strategy. *Nat Med* **19**, 488-493, doi:10.1038/nm.3092 (2013).

49 Kipp, K. R., Rezaei, M., Lin, L., Dewey, E. C. & Weimbs, T. A mild reduction of food intake slows disease progression in an orthologous mouse model of polycystic kidney disease. *Am J Physiol Renal Physiol* **310**, F726-F731, doi:10.1152/ajprenal.00551.2015 (2016).

50 Warner, G. *et al.* Food Restriction Ameliorates the Development of Polycystic Kidney Disease. *J Am Soc Nephrol* **27**, 1437-1447, doi:10.1681/ASN.2015020132 (2016).

51 Padovano, V., Podrini, C., Boletta, A. & Caplan, M. J. Metabolism and mitochondria in polycystic kidney disease research and therapy. *Nat Rev Nephrol* **14**, 678-687, doi:10.1038/s41581-018-0051-1 (2018).

52 Happe, H. *et al.* Toxic tubular injury in kidneys from Pkd1-deletion mice accelerates cystogenesis accompanied by dysregulated planar cell polarity and canonical Wnt signaling pathways. *Hum Mol Genet* **18**, 2532-2542, doi:10.1093/hmg/ddp190 (2009).

53 Weimbs, T. Polycystic kidney disease and renal injury repair: common pathways, fluid flow, and the function of polycystin-1. *Am J Physiol Renal Physiol* **293**, F1423-1432, doi:10.1152/ajprenal.00275.2007 (2007).

54 Kennefick, T. M. *et al.* Hypertension and renal injury in experimental polycystic kidney disease. *Kidney Int* **56**, 2181-2190, doi:10.1046/j.1523-1755.1999.00783.x (1999).

55 Patel, V. *et al.* Acute kidney injury and aberrant planar cell polarity induce cyst formation in mice lacking renal cilia. *Hum Mol Genet* **17**, 1578-1590, doi:10.1093/hmg/ddn045 (2008).

56 Low, S. H. *et al.* Polycystin-1, STAT6, and P100 function in a pathway that transduces ciliary mechanosensation and is activated in polycystic kidney disease. *Dev Cell* **10**, 57-69, doi:10.1016/j.devcel.2005.12.005 (2006).

57 Velculescu, V. E. *et al.* Characterization of the yeast transcriptome. *Cell* **88**, 243-251, doi:10.1016/s0092-8674(00)81845-0 (1997).

58 Velculescu, V. E., Zhang, L., Vogelstein, B. & Kinzler, K. W. Serial analysis of gene expression. *Science* **270**, 484-487, doi:10.1126/science.270.5235.484 (1995).

59 Heller, M. J. DNA microarray technology: devices, systems, and applications. *Annu Rev Biomed Eng* **4**, 129-153, doi:10.1146/annurev.bioeng.4.020702.153438 (2002).

60 Morozova, O., Hirst, M. & Marra, M. A. Applications of new sequencing technologies for transcriptome analysis. *Annu Rev Genomics Hum Genet* **10**, 135-151, doi:10.1146/annurev-genom-082908-145957 (2009).

61 Bainbridge, M. N. *et al.* Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. *BMC Genomics* **7**, 246, doi:10.1186/1471-2164-7-246 (2006).

62 Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**, 621-628, doi:10.1038/nmeth.1226 (2008).

63 Wilhelm, B. T. *et al.* Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* **453**, 1239-1243, doi:10.1038/nature07002 (2008).

64 Hangauer, M. J., Vaughn, I. W. & McManus, M. T. Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. *PLoS Genet* **9**, e1003569, doi:10.1371/journal.pgen.1003569 (2013).

65 Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**, 511-515, doi:10.1038/nbt.1621 (2010).

66 Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455-461, doi:10.1038/nature12787 (2014).

67 Kim, T. K. *et al.* Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**, 182-187, doi:10.1038/nature09033 (2010).

68 Camarena, L., Bruno, V., Euskirchen, G., Poggio, S. & Snyder, M. Molecular mechanisms of ethanol-induced pathogenesis revealed by RNA-sequencing. *PLoS Pathog* **6**, e1000834, doi:10.1371/journal.ppat.1000834 (2010).

69 Wang, E. T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470-476, doi:10.1038/nature07509 (2008).

70 Roberts, A., Trapnell, C., Donaghey, J., Rinn, J. L. & Pachter, L. Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biol* **12**, R22, doi:10.1186/gb-2011-12-3-r22 (2011).

71 Tarazona, S., Garcia-Alcalde, F., Dopazo, J., Ferrer, A. & Conesa, A. Differential expression in RNA-seq: a matter of depth. *Genome Res* **21**, 2213-2223, doi:10.1101/gr.124321.111 (2011).

72 Costa, V., Aprile, M., Esposito, R. & Ciccodicola, A. RNA-Seq and human complex diseases: recent accomplishments and future perspectives. *Eur J Hum Genet* **21**, 134-142, doi:10.1038/ejhg.2012.129 (2013).

73 Khurana, E. *et al.* Role of non-coding sequence variants in cancer. *Nat Rev Genet* **17**, 93-108, doi:10.1038/nrg.2015.17 (2016).

74 Lim, W. K. *et al.* Exome sequencing identifies highly recurrent MED12 somatic mutations in breast fibroadenoma. *Nat Genet* **46**, 877-880, doi:10.1038/ng.3037 (2014).

75 Majewski, J. & Pastinen, T. The study of eQTL variations by RNA-seq: from SNPs to phenotypes. *Trends Genet* **27**, 72-79, doi:10.1016/j.tig.2010.10.006 (2011).

76 Yang, Y. *et al.* Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. *Nat Commun* **5**, 3231, doi:10.1038/ncomms4231 (2014).

77 Voineagu, I. *et al.* Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* **474**, 380-384, doi:10.1038/nature10110 (2011).

78 Wacker, S. A., Houghtaling, B. R., Elemento, O. & Kapoor, T. M. Using transcriptome sequencing to identify mechanisms of drug action and resistance. *Nat Chem Biol* **8**, 235-237, doi:10.1038/nchembio.779 (2012).

79 Cui, Y. & Paules, R. S. Use of transcriptomics in understanding mechanisms of drug-induced toxicity. *Pharmacogenomics* **11**, 573-585, doi:10.2217/pgs.10.37 (2010).

80 Dweep, H., Sticht, C., Kharkar, A., Pandey, P. & Gretz, N. Parallel analysis of mRNA and microRNA microarray profiles to explore functional regulatory patterns in polycystic kidney disease: using PKD/Mhm rat model. *PLoS One* **8**, e53780, doi:10.1371/journal.pone.0053780 (2013).

81 Kugita, M. *et al.* Global gene expression profiling in early-stage polycystic kidney disease in the Han:SPRD Cy rat identifies a role for RXR signaling. *Am J Physiol Renal Physiol* **300**, F177-188, doi:10.1152/ajprenal.00470.2010 (2011).

82 Song, X. *et al.* Systems biology of autosomal dominant polycystic kidney disease (ADPKD): computational identification of gene expression pathways and integrated regulatory networks. *Hum Mol Genet* **18**, 2328-2343, doi:10.1093/hmg/ddp165 (2009).

83 Wallace, D. P. *et al.* Periostin induces proliferation of human autosomal dominant polycystic kidney cells through alphaV-integrin receptor. *Am J Physiol Renal Physiol* **295**, F1463-1471, doi:10.1152/ajprenal.90266.2008 (2008).

84 Riera, M., Burtey, S. & Fontes, M. Transcriptome analysis of a rat PKD model: Importance of genes involved in extracellular matrix metabolism. *Kidney Int* **69**, 1558-1563, doi:10.1038/sj.ki.5000309 (2006).

85 Schieren, G. *et al.* Gene profiling of polycystic kidneys. *Nephrol Dial Transplant* **21**, 1816-1824, doi:10.1093/ndt/gfl071 (2006).

86 Allen, E. *et al.* Loss of polycystin-1 or polycystin-2 results in dysregulated apolipoprotein expression in murine tissues via alterations in nuclear hormone receptors. *Hum Mol Genet* **15**, 11-21, doi:10.1093/hmg/ddi421 (2006).

87 Husson, H. *et al.* New insights into ADPKD molecular pathways using combination of SAGE and microarray technologies. *Genomics* **84**, 497-510, doi:10.1016/j.ygeno.2004.03.009 (2004).

88 Joly, D. *et al.* Beta4 integrin and laminin 5 are aberrantly expressed in polycystic kidney disease: role in increased cell adhesion and migration. *Am J Pathol* **163**, 1791-1800, doi:10.1016/s0002-9440(10)63539-0 (2003).

89 Menezes, L. F. *et al.* Network analysis of a Pkd1-mouse model of autosomal dominant polycystic kidney disease identifies HNF4alpha as a disease modifier. *PLoS Genet* **8**, e1003053, doi:10.1371/journal.pgen.1003053 (2012).

90 Pandey, P., Qin, S., Ho, J., Zhou, J. & Kreidberg, J. A. Systems biology approach to identify transcriptome reprogramming and candidate microRNA targets during the progression of polycystic kidney disease. *BMC Syst Biol* **5**, 56, doi:10.1186/1752-0509-5-56 (2011).

91 Stephens, Z. D. *et al.* Big Data: Astronomical or Genomical? *PLoS Biol* **13**, e1002195, doi:10.1371/journal.pbio.1002195 (2015).

92 Hsi-Yang Fritz, M., Leinonen, R., Cochrane, G. & Birney, E. Efficient storage of high throughput DNA sequencing data using reference-based compression. *Genome Res* **21**, 734-740, doi:10.1101/gr.114819.110 (2011).

93 Schadt, E. E., Linderman, M. D., Sorenson, J., Lee, L. & Nolan, G. P. Computational solutions to large-scale data management and analysis. *Nat Rev Genet* **11**, 647-657, doi:10.1038/nrg2857 (2010).

94 Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018, doi:10.1038/sdata.2016.18 (2016).

95 Stevens, R., Bodenreider, O. & Lussier, Y. A. Semantic webs for life sciences. *Pac Symp Biocomput*, 112-115 (2006).
96 Hettne, K. M. *et al.* The Implicitome: A Resource for Rationalizing Gene-Disease Associations. *PLoS One* **11**, e0149621, doi:10.1371/journal.pone.0149621 (2016).
97 Chen, B., Ding, Y. & Wild, D. J. Assessing drug target association using semantic linked data. *PLoS Comput Biol* **8**, e1002574, doi:10.1371/journal.pcbi.1002574 (2012).

CHAPTER 2
META-ANALYSIS OF POLYCYSTIC KIDNEY DISEASE EXPRESSION PROFILES DEFINES STRONG INVOLVEMENT OF INJURY REPAIR PROCESSES

Tareq B. Malas*; Chiara Formica*; Wouter N. Leonhard; Pooja Rao; Zoraide Granchi; Marco Roos; Dorien J.M. Peters; Peter A.C. 't Hoen

American Journal of Physiology-Renal Physiology,312,4,F806-F817,2017, American Physiological Society Bethesda, MD
January 2017

*Contributed equally

Meta-analysis of polycystic kidney disease expression profiles defines strong involvement of injury repair processes

Tareq B. Malas,[1]* Chiara Formica,[1]* Wouter N. Leonhard,[1] Pooja Rao,[2] Zoraide Granchi,[2] Marco Roos,[1] Dorien J. M. Peters,[1] and Peter A. C. 't Hoen[1]

[1]*Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands; and* [2]*GenomeScan, Leiden, The Netherlands*

Submitted 16 December 2016; accepted in final form 30 January 2017

Malas TB, Formica C, Leonhard WN, Rao P, Granchi Z, Roos M, Peters DJ, 't Hoen PA. Meta-analysis of polycystic kidney disease expression profiles defines strong involvement of injury repair processes. *Am J Physiol Renal Physiol* 312: F806–F817, 2017. First published February 1, 2017; doi:10.1152/ajprenal.00653.2016.—Polycystic kidney disease (PKD) is a major cause of end-stage renal disease. The disease mechanisms are not well understood and the pathogenesis toward renal failure remains elusive. In this study, we present the first RNASeq analysis of a *Pkd1*-mutant mouse model in a combined meta-analysis with other published PKD expression profiles. We introduce the PKD Signature, a set of 1,515 genes that are commonly dysregulated in PKD studies. We show that the signature genes include many known and novel PKD-related genes and functions. Moreover, genes with a role in injury repair, as evidenced by expression data and/or automated literature analysis, were significantly enriched in the PKD Signature, with 35% of the PKD Signature genes being directly implicated in injury repair. NF-κB signaling, epithelial-mesenchymal transition, inflammatory response, hypoxia, and metabolism were among the most prominent injury or repair-related biological processes with a role in the PKD etiology. Novel PKD genes with a role in PKD and in injury were confirmed in another *Pkd1*-mutant mouse model as well as in animals treated with a nephrotoxic agent. We propose that compounds that can modulate the injury-repair response could be valuable drug candidates for PKD treatment.

polycystic kidney disease; injury repair; Meta-analysis; RNASeq; expression signature; literature mining

POLYCYSTIC KIDNEY DISEASE (PKD) is a genetic disease of the kidney characterized by the gradual replacement of normal kidney parenchyma by fluid-filled cysts and fibrotic tissue. Autosomal dominant polycystic kidney disease (ADPKD) is caused by mutations in the *PKD1* or *PKD2* gene and the less frequent autosomal recessive form, autosomal recessive polycystic kidney disease (ARPKD), is caused by mutations in the *PKHD1* gene (4, 40, 67). It is not entirely clear why the disruption of these genes lead to PKD and what functions their protein products might have in normal and diseased kidneys. Furthermore, it is expected that a vasopressin V2 receptor antagonist, recently approved in Europe, will probably not be sufficient for life-long treatment, warranting the search for additional therapies (65). Therefore, a detailed knowledge of the molecular pathology and signaling pathways at different phases of the disease is needed. To answer these questions, several PKD-related expression profiling studies have been conducted in the last decade (10, 20, 30, 39, 43–45, 51, 54, 56). These studies, however, varied considerably by the type and number of the samples used, i.e., cystic and normal cell lines, patient-derived material, kidneys from different rat or mouse models at different stages of the disease, and the analysis platforms and methods, as reviewed by Menezes and Germino (37). As a result, a variety of different pathways and processes were suggested to contribute to the disease but with no strong evidence of which of these differences in the reported results are a consequence of experimental biases or disease complexity.

Several studies indicated a tissue injury repair component in the pathology of PKD (16, 24, 70). Indeed, there are numerous similarities between PKD and renal injury, since both phenotypes are accompanied by a combination of processes including proliferation, secretion of growth factors, as well as inflammation. Weimbs (70) proposed a model where *Polycystin-1* (*PC1*), the protein encoded by *PKD1*, and primary cilia have a critical function in sensing renal injury by detecting changes in luminal fluid flow and triggering proliferation. Besides a proposed mechanistic overlap, several studies showed that renal injury could stimulate cyst progression. For example, nephrotoxic injury in an ADPKD adult mouse model resulted in accelerated cyst formation and a more progressive phenotype (16). This is further supported by findings that ischemic reperfusion injury and also tubular cell hypertrophy following unilateral nephrectomy accelerated PKD (6, 16, 33, 47, 62). Although the link between PKD and renal injury seems rather strong, until now a thorough comparison between the two conditions at the molecular level has not been made, and little is known about the key genes and pathways shared between the two.

In this work, we performed a meta-analysis of PKD expression profiles to come to a consistent expression signature of the disease that minimizes experimental and technology biases. We also performed an in-depth comparison between PKD and injury repair models to characterize genes and functions involved in injury repair processes and PKD pathology. The novelty of our approach lies in its ability to overcome single study biases in describing the PKD and injury repair expression signatures, and in the combined use of experimental data and prior knowledge, retrieved from databases and mined from the literature.

* T. B. Malas and C. Formica contributed equally to this work.

Address for reprint requests and other correspondence: P. C. 't Hoen, Dept. of Human Genetics, Leiden Univ. Medical Center, Einthovenweg 20, 2333 ZC Leiden, The Netherlands (e-mail: p.a.c.hoen@lumc.nl).

MATERIALS AND METHODS

Experimental Animals and RNA Sequencing

Pkd1 mutant and wild-type mice. The inducible kidney-specific *Pkd1*-deletion mouse model (tam-KspCad-CreER^{T2};*Pkd1*$^{lox2\text{-}11/lox2\text{-}11}$, referred to as iKsp-*Pkd1*^{del}) and tamoxifen treatments were previously described (32, 42). RNA sequencing was done on five wild-type (WT) mice and kidneys of four iKsp-*Pkd1*^{del} mice with gene disruption at the age of 38-40 days (mutant). Mutant mice, euthanized 84 days later, had moderate cystic disease. The local animal experimental committee of the Leiden University Medical Center and the Commission Biotechnology in Animals of the Dutch Ministry of Agriculture approved the experiments performed.

DCVC injury model. WT mice were fed with tamoxifen (5 mg/day, 3 consecutive days) at adult age, i.e. between 13 to 14 wk of age (42) as a control for the tamoxifen treatment used in the iKsp-*Pkd1*^{del} mice. Renal injury was induced 1 wk after tamoxifen administration by a single intraperitoneal injection of *S*-(1,2-dichlorovinyl)-L-cysteine (DCVC) (15 mg/kg). Mice were euthanized at determined time points (1, 2, 5, 10, and 24 wk after DCVC injection). RNA sequencing was performed on the DCVC-injected WT mice euthanized at 1, 2, and 5 wk after DCVC (4 mice per each time point).

RNA sequencing methodology. RNA sequencing was performed on the Illumina Hi Seq 2500. mRNA-Seq Sample Prep Kit was used to process the samples according to the manufacturer's protocol. Briefly, mRNA was isolated from total RNA using the oligo-dT magnetic beads. After fragmentation of the mRNA, a cDNA synthesis was performed. This was used for ligation with the sequencing adapters and PCR amplification of the resulting product. The quality and yield after sample preparation were measured with a DNA 1000 Lab-on-a-Chip. The expected broad peak between 300 and 500 bp was observed.

Clustering and DNA sequencing using the Illumina cBot and HiSeq 2500 was performed according to manufacturer's protocols. A concentration of 15.0 pM of DNA was used.

HiSeq control software HCS v2.2.38 was used. Image analysis, base calling, and quality check were performed with the Illumina data analysis pipeline RTA v1.18.64 and Bcl2fastq v1.8.4. All samples had a quality score Q30 for more than 93.6% of reads.

Resulting reads were aligned to the mouse reference genome version GRCm38 (68) using Tophat2 (25) followed by bowtie2 (31) in the local highly sensitive mode (bowtie2-local-very-sensitive-local). After alignment, HTSeq-count (3) (Version 0.6.1) was used to estimate gene expression by counting reads that were mapped to each gene using default options. Differential gene expression analysis was performed using *limma* package with default parameters after applying Voom transformation (52). Differentially expressed genes were selected where false discovery rate (FDR) is <0.05. Data were deposited in ArrayExpress (27) and given the following identifier E-MTAB-5319.

Experimental Animals and Fluidigm Assay

Pkd1 mutant, WT, and DCVC-induced mice. WT mice and iKsp-*Pkd1*^{del} mice were fed with tamoxifen (5 mg/day, 3 consecutive days) in adult mice, i.e., between 13 and 14 wk of age (42), to achieve *Pkd1* gene inactivation in the mutant. Renal injury was induced in WT mice by a single intraperitoneal injection of DCVC (15 mg/kg). All mice were euthanized at determined time points (1,2, 5, 10, and 24 wk after DCVC injection for the DCVC-induced WTs and respective time points for the non-DCVC-treated mice).

Fluidigm quantitative PCR and data processing. The TaqMan Gene Expression Assays of the selected genes and three housekeeping genes were obtained from Applied Biosystems. Best coverage probes were used, according to the sample characteristics. Real-Time PCR analysis was performed at GenomeScan (GenomeScan, Leiden, The Netherlands) using the 96.96 BioMark Dynamic Array for Real-Time PCR (Fluidigm, San Francisco, CA), according to the manufacturer's instructions. Before use on the BioMark array, the cDNA was first subjected to 14 cycles of Specific Target Amplification using a 0.2× mixture of all Taqman Gene Expression assays in combination with the TaqMan PreAmp Master Mix (Applied Biosystems), followed by fivefold dilution. Thermal cycling and real-time imaging of the BioMark array was done on the BioMark instrument, using the default Taqman PCR protocol with an annealing temperature of 60°C and a total of 35 cycles of PCR. Ct values (cycle threshold values) were extracted using the BioMark Real-Time PCR analysis software (version 3.0.2) and the threshold default value of 0.65. The quality of the amplification curves was checked for each reaction, evaluating the curve shape and signal level. For each gene, Ct values were normalized based on the geometric mean of the housekeeping genes (Rplp0, Hnrnpa2b1, and Ywhaz) and then compared across the samples for differential expression in PKD (PKD vs. WT) and injury (WT injury induced by DCVC vs. WT) by using ANOVA. A $P < 0.01$ cut-off was used to determine significantly dysregulated genes.

Data Acquisition and Meta-analysis

Meta-Analysis PKD Signature. In addition to our iKsp-*Pkd1*^{del}, public expression profiling experiments of PKD were downloaded from ArrayExpress using ArrayExpress R Package (23) available on Bioconductor (15). Published PKD expression profiling studies were included based on the following selection criteria: *1*) expansion renal cystic tissues were obtained from animal models with disruption of a gene either involved in ADPKD or in ARPKD (PCK rat used as a model for PKD) or taken from ADPKD patients. For the large phenotypic and physiological differences between postnatal and embryonic PKD models, embryonic PKD models were excluded. Asymptotic models were excluded as well. *2*) The study included at least three biological and technical replicates for mutant and control tissues (WTs). *3*) Normalized gene expression values were publicly available for all samples. Processed data were used for each data set and log transformed it if data were provided on a linear scale. Then, we calculated differentially expressed genes for each study using the *limma* package with default parameters (52). Genes were considered significantly dysregulated if they had FDR <0.05 and <0.0005 for the validation study. When a study has different models or phenotypes of the disease, we processed the samples independently when calculating the differentially expressed genes, and only included in our analysis the models/samples that are useful to us (refer to the relevant table for sample description). When including more than one model from a single study, we took the resulting lists of differentially regulated genes of each model and combined them in a one gene list per study. Conversion to human homologs was done by using the *db2db* tool part of BioDBnet (41).

Public injury models and other kidney diseases. In addition to our DCVC-treated WTs, we included six published studies of renal ischemia. Study inclusion criteria were based on the availability of treated and nontreated WTs and availability of data as described in the previous paragraph. Other kidney diseases were included as mentioned in RESULTS. Differentially expressed genes for each study were calculated as described in the previous paragraph.

Literature-Based Signatures

Literature-based signatures were obtained using Biosemantics Concept Profile technology (17a, 21), calculating the literature association scores between all *Homo sapiens* genes with concept profiles and the literature concept profiles of renal injury repair in PubMed until July 2012. The literature association (concept profile matching) score is based on the strength of explicit associations (the 2 concepts co-occurring in the same PubMed abstract) and

implicit associations (the 2 concepts not co-occurring in the same abstract together, but each in abstracts with the same 3rd concept). The literature mining technology handles and disambiguates gene names and symbols.

Functional Enrichment Analysis

Functional enrichment analysis was performed against Molecular Signature Database (MSigDB) collections (58, 59) using standard hypergeometric distribution with correction for multiple hypotheses testing using the FDR. Annotation with DAVID v6.8[27,28] was performed with default parameters for confirmation purposes where mentioned in the text.

The significance of the overlaps between gene profiles and gene sets, were evaluated with a χ^2-test ($P < 0.05$) and the representation factor (RF) calculated as follows: x = # of genes in common between two groups, n = # of genes in group 1, D = # of genes in group 2, N = total genes available, the RF = x/expected # of genes, expected # of genes = $(n \times D)/N$.

RESULTS

Identification of PKD Signature

To identify a comprehensive expression signature of PKD during disease progression we generated new RNASeq-based expression profiles of inducible, kidney epithelium-specific *Pkd1*-deletion mice (iKsp-*Pkd1*^{del}) with moderate cystic disease and matched controls (32, 42) (Fig. 1*A*). We identified 2,376 genes (*FDR* < 0.05) that clearly distinguished iKsp-*Pkd1*^{del} from WT mice (Data Set1, Fig. 2*A*). Next, we compared our expression profile to other publicly available, PKD expression studies (Fig. 1*A*). We used stringent study inclusion criteria (see METHODS) and identified three studies suitable for meta-analysis (39, 43, 56) (Table 1). By comparing the independent PKD expression studies (our data set included), we assessed the level of similarity/dissimilarity between each study and defined a group of high confidence genes that are consistently dysregulated in PKD. Every significantly dysregulated gene (PKD vs. WT) obtained from the studies, was classified as a shared gene (if dysregulated in at least 2 independent studies) or a unique gene (if dysregulated in just a single study). Strikingly, only ~22% (N = 1515) of all dysregulated genes from all PKD studies (N = 6,963) were shared, of which the vast majority (~86% of the 1,515 genes) were dysregulated in just two studies (Fig. 2*B*). Moreover, none of the PKD studies used in this analysis had more than 50% of its dysregulated genes shared with any of the other three studies (Fig. 2*C*).

To arrive to a robust PKD Signature, we selected PKD genes that were significantly dysregulated in at least two independent studies (50% of the studies). Thus our PKD Signature consists

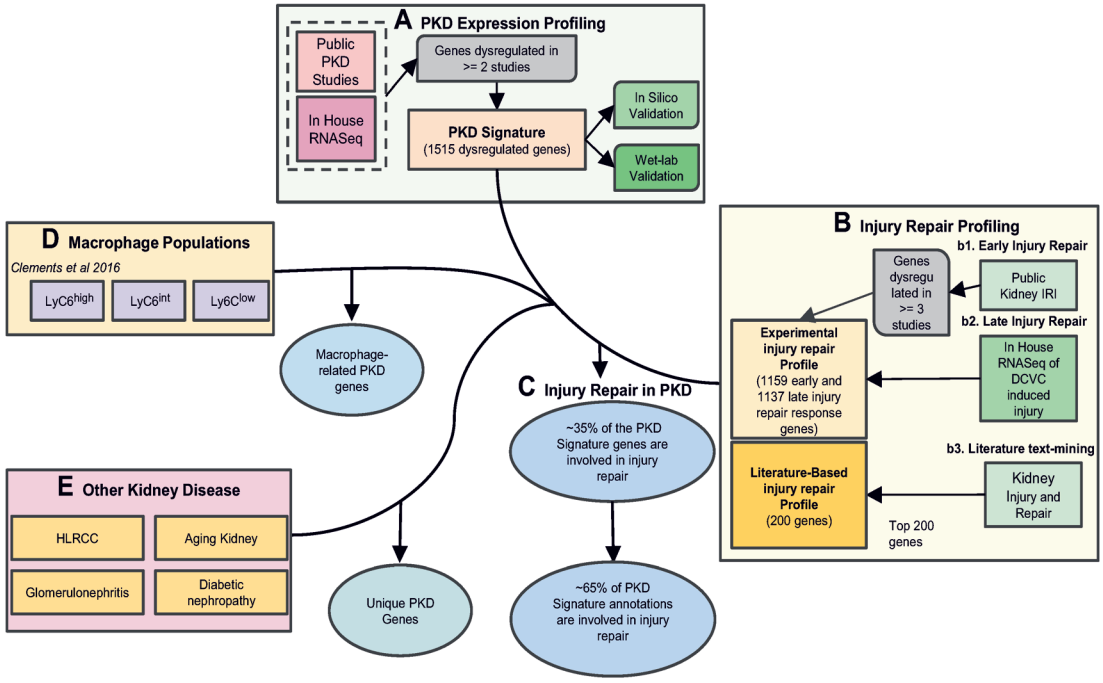

Fig. 1. Overview of the approach used to identify the polycystic kidney disease (PKD) Signature and comparison to renal injury and repair, macrophages and other kidney diseases. The approach consisted of 5 steps. *A*: the PKD Signature was defined by combining publicly available PKD expression profiling studies with our in-house RNAseq of iKsp-*Pkd1*^{del} in mice. *B*: the Injury Repair Profile was defined by experimental expression profiles of kidneys with ischemia-reperfusion injury (IRI) (*b1*), in-house RNAseq of kidneys from *S*-(1,2-dichlorovinyl)-L-cysteine (DCVC)-treated animals (*b2*) and literature-based text-mining of genes associated with injury terms in PubMed abstracts (*b3*). *C*: comparing significantly dysregulated genes from PKD Signature and Injury Repair Profile we identified the injury repair component of PKD, which consists of ~35% of the genes implicated in PKD. *D*: we used the data produced by Clements et. al (11) in 2016 of the different macrophage populations triggered after renal injury to identify macrophage-related genes in PKD. *E*: we acquired expression profiling experiments of different renal diseases and compared them with the PKD Signature to identify the overlapping genes and the unique PKD Signature genes. HLRCC, hereditary leiomyomatosis and renal cell cancer.

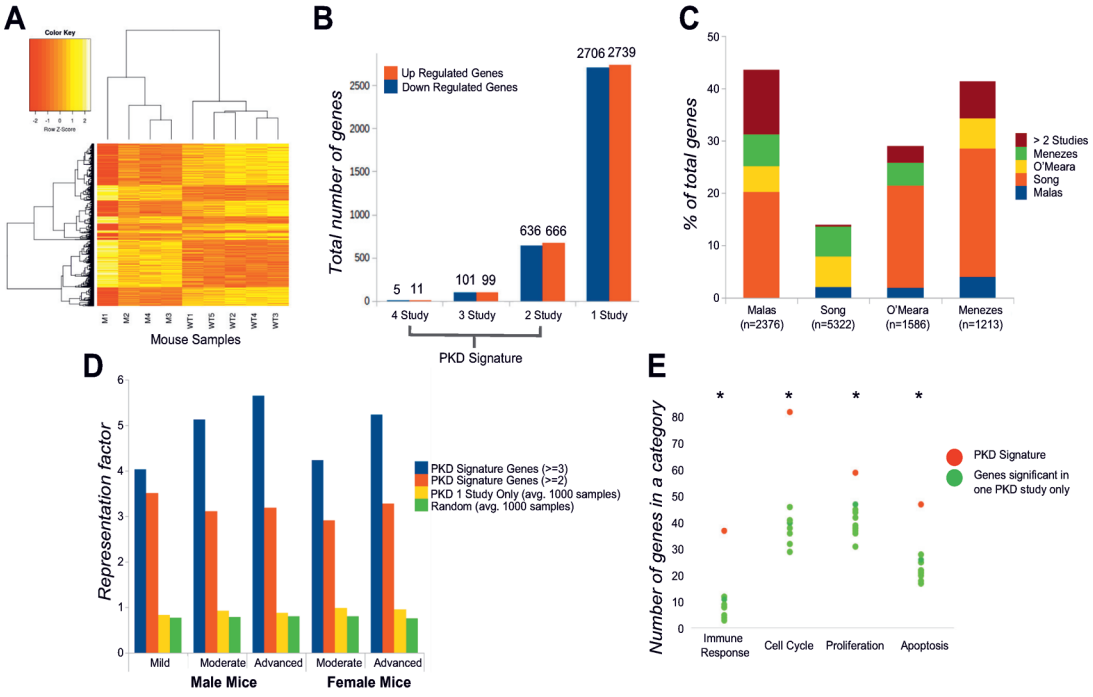

Fig. 2. Identification of the PKD Signature. *A*: heat map showing the expression values of all differentially expressed genes in iKsp-*Pkd1*^{del} (M) compared with wild-type mice (WT). Expression values were normalized using the Voom function in *limma* R Package. Hierarchical clustering was applied on the samples and values were scaled by row. *B*: analysis of the number of overlapping genes from the 4 studies included in the meta-analysis. *C*: comparison of the significantly dysregulated genes in each of the 4 studies revealed that none of the studies shared more than 50% of its genes with other studies. *D*: bar chart plotting the enrichment (representation factor, *y*-axis) of the PKD Signature (blue and red bars), the genes dysregulated only in a single PKD study (yellow) and a random set of genes (green) in an independent data set from *Pkd1cko* mice at mild, moderate, and advanced stages of the disease, in males and females. *E*: a dot plot showing the number of genes from the PKD Signature (red) and from 10 equally sized gene-sets randomly sampled from the genes dysregulated in only one study (green) with functional annotations previously attributed to PKD. *$P < 0.05$ (binomial test, where the fraction of genes in the PKD Signature was compared with the expected fraction of genes in a functional category (averaged gene counts from the 10 random sets). Functional categories were based on 3 annotation databases (GSEA_HALLMARKS, KEGG, and Gene Ontology–*Biological Processes*).

of 1,515 genes (1,641 mouse homologs), comprising 775 up- and 740 downregulated genes (Data Set2a and 2b).

Validation of the PKD Signature in an Independent Data Set

We validated the PKD Signature in an independent data set that was published during the writing of this manuscript (38). This data set includes 80 mouse samples with induced deletion of *Pkd1* at P40 (14 control females, 21 mutant females, 19 control males, and 26 mutant males). We evaluated the overlap by calculating the representation factor (RF), defined as the number of overlapping genes divided by the expected number of overlapping genes drawn from two independent groups. We looked for the enrichment of the PKD Signature genes in the genes differentially expressed (FDR < 0.0005) in mild, moderate, and advanced stages of the disease in males and in females (Fig. 2*D*). We repeated the enrichment test with equally sized gene sets from the genes that are dysregulated in only one PKD study and from randomly sampled genes. The results revealed that the PKD Signature is at least three times more enriched in the validation data set in contrast to the random genes and genes dysregulated in only one study. This significance is observed in all stages of the PKD disease and in both genders.

Table 1. *Expression profiling studies used in the definition of the PKD Signature*

Authors	Organism/Sex	Accession Number	Data Sets Included	No. of DEGs after HC	No. of DEGs in PKD Signature
Menezes et al. (39)	Mouse/M and F	GSE32586	*Pkd1cko.P12* vs. WT.12	40	17
			Pkd1cko.P14 vs. WT.14	1,200	594
O'Meara et al. (43)	Rat/M	GSE33056	PCK vs. SD	1,586	561
Song et al. (56)	Human/NM	GSE7869	Noncystic vs. normal	280	68
			Small cysts vs. normal	4,554	1,025
			Medium cysts vs. normal	4,650	1,033
Malas et al. (this study)	Mouse/M	E-MTAB-5319	*Pkd1cko.P40* vs.WT	2,376	924

PKD, polycystic kidney disease; DEGs, differentially expressed genes; HC, homolog conversion; SD, Sprague-Dawley; WT, wild type; M, males; F, females; NM, not mentioned in the METHODS of the paper.

Functional Annotation of the PKD Signature

We annotated the up- and downregulated gene sets using the Molecular Signature Database Hallmarks (MSigDB) (34) and The Database for Annotation, Visualization, and Integrated Discovery (DAVID) v6.8 (18, 19) (Data Set3). Both resources revealed the strong downregulation of mitochondrial and peroxisome genes, specifically those involved in fatty acid metabolism, lipogenesis, and oxidation-reduction process. In addition, we see strong dysregulation of genes involved in ion transport (i.e., *CP*, *SLC4A1*, *SLC2A9*, and *SLC12A1*).

On the other hand, several pathways and processes are upregulated in the PKD Signature listed in Data Set3. We further grouped the genes shared between three or more studies into defined gene families (protein kinases, cytokines and growth factors, transcription factors, and oncogenes) to facilitate their usage in translational research (Data Set3c and 3d).

To measure the relevance of the PKD Signature at the molecular-function level, we defined four categories that are known to be dysregulated in PKD: cell cycle (46), apoptosis (13), proliferation (71), and inflammatory response (61). We found that the PKD Signature had significantly more functional annotations related to these categories than equally sized random sets of genes pooled from the genes significant in only one PKD study (binomial test, $P < 0.0005$) (Fig. 2*E*).

Identification of Injury Repair Genes in the PKD Signature

Experimental injury repair profile. We acquired data sets from renal ischemia reperfusion (IRI) experiments in murine models and identified six studies (3 mouse and 3 rat) that met our inclusion criteria (Table 2) (Fig. 1*B*) (9, 12, 28, 35, 57, 73). Differentially expressed genes from all of the six renal injury repair studies were enriched within the PKD Signature (RF $>$ 1 and $P < 0.05$) (Fig. 3*B*). The studies varied in the reperfusion time, from immediately harvesting the samples after reperfusion to waiting for up to 120 h. Using these samples, we defined a 1,193-gene signature of early injury repair response by combining the six studies and looking for consistently dysregulated genes in at least 50% of the studies (Data Set4a).

Several renal IRI studies reported that the maximum peak of renal damage is reached at the 3rd day of injury induction and gradually drops as the kidney repairs itself, until about *day 7* when the tissue looks mostly repaired. Since none of the IRI studies we could include publicly harvested samples within the late injury period (after 7 days), we generated novel data by inducing renal injury in WT mice using the nephrotoxic DCVC. The samples were harvested at 1, 2, and 5 wk after injury induction. Measurements of blood urea showed that mice have normal functional kidneys at 7 days after injury, confirming that most of the acute injury insult is repaired at that time point (Fig. 3*A*). Renal histology sections at 1 wk show slightly dilated tubules with protein casts, infiltrating cells, and fibrosis at the cortico-medullary junction, the most damage-sensitive region of the kidney. At 2 wk after DCVC treatment the tissue morphology is improved and at 5 wk, and the kidneys have almost completely been recovered from the injury (Fig. 3*A*). Comparing the different injury repair time points from our injury-induced model against each other we identified 1,137 differentially expressed genes (Data Set4b). Comparing the functional annotations of the early and late injury repair genes showed that they are distinct (Fig. 3*C*). We combined the early and late injury repair response genes to establish a general Experimental Injury Repair Profile, consisting of 2137 genes.

Table 2. *Published expression profiling studies used in the definition of the experimental injury repair profile*

Authors	Organism	Accession	Injury Model	Time after Reperfusion
Chen et al. (9)	Mouse	GSE34351	ARI	4 h
Correa-Costa et al. (12)	Mouse	GSE39548	ARI	6 h
Liu et al. (35)	Mouse	GSE52004	ARI	24 h
Yuen et al. (73)	Rat	GSE3219	ARI	2, 8 h
Krishnamoorthy et al. (28)	Rat	GSE27274	ARI	6, 24, 120 h
Speir et al. (57)	Rat	GSE58438	ARI	Immediate

ARI, acute renal injury.

Literature injury repair profile. The Literature Injury Repair Profile is based on published knowledge of genes involved in injury repair processes in PubMed. We set our text-mining algorithm to retrieve the top 200 genes that are most associated with renal injury in PubMed abstracts (Data Set4c). In addition, as negative controls, we retrieved the top 200 genes most associated with other conditions that we believe to be unrelated to PKD, namely fertility and Parkinson's disease. The literature renal injury repair genes were significantly enriched in the PKD Signature (RF = 2), contrary to the two negative sets (fertility RF = 0.9, Parkinson's disease RF = 1.1).

Injury Repair Processes in PKD

To find injury repair involvement in PKD, we compared the PKD Signature with the Experimental Injury Repair Profile and the Literature Injury Repair Profile (Fig. 1*C*). Our analysis revealed that both injury repair profiles were significantly enriched in the PKD Signature (Fig. 3*D*), with *P* values of 1.3×10^{-29} and 2.8×10^{-7} for the Experimental and Literature Injury Repair Profiles, respectively. Of the total PKD Signature genes, 35% (581 genes) are involved in injury repair processes (Data Set5).

We extended the functional analysis that we performed on the PKD Signature to the injury repair and noninjury repair genes in PKD. For each set of genes, we identified the most enriched functional terms from MSigDB with stringent FDR cutoff ($<$ 1e-11). Interestingly, 22 terms were found to be more enriched in the injury repair genes and only 1 term is more enriched in the noninjury repair genes when compared with each other (Fig. 3*E*, Data Set6). This demonstrates the cohesiveness of the injury repair genes. Many of the injury repair functions are related to NF-kB signaling, epithelial-mesenchymal transition, inflammatory response, hypoxia, and metabolism. Although these functional terms are expected in a well-defined injury repair signature, they demonstrate that we zoomed in on a relevant group of genes (Table 3).

Injury repair in Different PKD Disease Stages

We evaluated the enrichment of the genes in the signature at the different phases of disease progression. We used the data of Menezes et al. (26) that included mouse samples from mild, moderate, and advanced PKD disease stages. Injury repair and noninjury repair PKD genes are enriched in all stages of the disease (RF $>$ 1) (Fig. 4*A*). However, in the mild phase of the

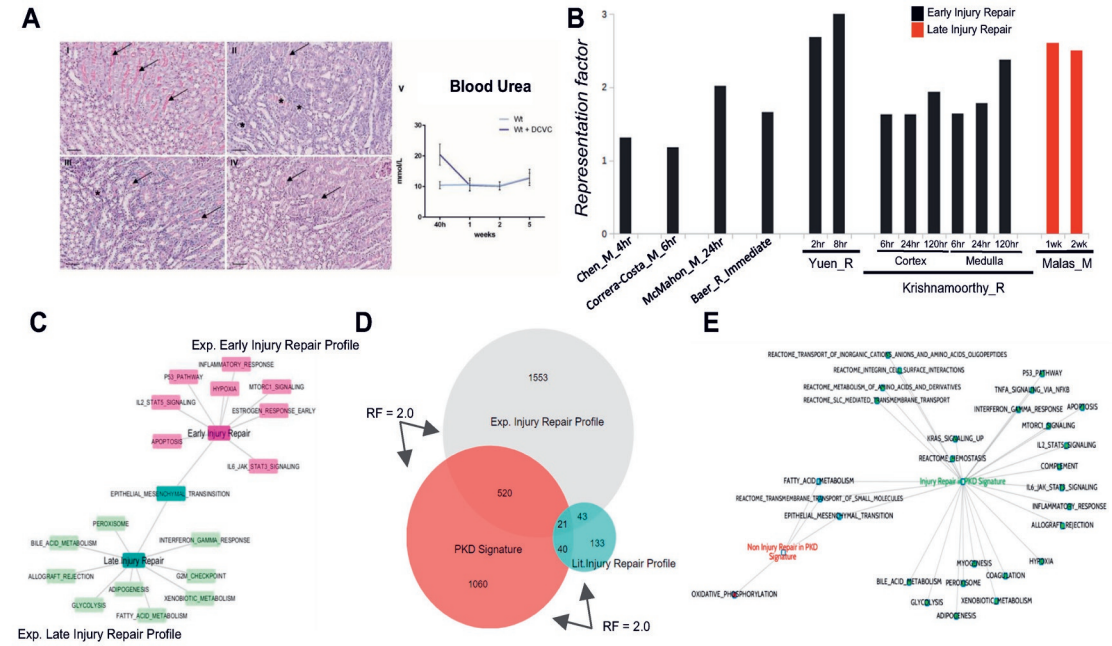

Fig. 3. Identification of the injury repair component of the PKD Signature. *A*: injury repair progression in WT mice injected with DCVC. *I–IV*: periodic-acid Schiff (PAS) staining on formalin fixed, paraffin embedded kidney sections; scale bar = 50 um. *Dilated tubules with luminal protein inclusions; →normal brush border. *I*: WT without injury. Normal morphology with visible brush border. *II*: 1 wk after DCVC injection damaged and slightly dilated tubules with luminal protein inclusions were observed. *III*: 2 wk after DCVC the tissue is almost repaired with only sporadic protein inclusion observable but some infiltrating cells present. *IV*: 5 wk after DCVC injection the tissue was almost completely repaired and less infiltrating cells were present. *V*: blood urea (BU) levels (mmol/l) at different time points after DCVC injection, peaking at 40 h and returning to baseline after 2 wk, when the tissue is repaired. *B*: enrichment (representation factor, *y*-axis) of each injury data set in the PKD Signature. Data set from mice (M) or rats (R) with early injury response (IRI, black) and late injury repair response (DCVC treatment, red) were all enriched in the PKD Signature. *C*: the top 10 enriched terms for the early injury repair (pink) and late injury repair (green) genes based on MSigDB Hallmark categories [false discovery rate (FDR) < 0.05]. On the one hand, the early injury repair genes had apoptosis, hypoxia and inflammatory response related pathways in the most enriched functions identified from MSigDB and, on the second hand, the late injury repair genes had functions related to peroxisome, glycolysis and fatty acid metabolism amongst the most enriched functions. *D*: comparison of the experimental (Exp.) and literature (Lit.) based injury repair profiles and their overlap with the PKD Signature. Approximately, 35% (581/1,641) of the PKD Signature genes are involved in injury repair related functions. Representation factor (RF) denotes the level of enrichment of each injury repair profile in the PKD Signature (RF of Experimental Injury Repair = 2.0, RF of Literature Injury Repair = 2.0) where values greater than 1 denote strong enrichment. *E*: a clear distinction in the functional profiles of the injury (red) and noninjury (green) components of the PKD Signature, as represented in a network view of the more enriched terms from MSigDB (FDR < 1e-11) for each of the 2 components.

disease, the noninjury repair is twice as enriched in comparison to the injury repair genes. In the advanced stage of the disease, the opposite is observed. These results are in accordance with our understanding of the disease, as more injury repair processes are expected with disease progression. That said, injury repair genes are enriched in the early phases of PKD and appear to be involved in cell cycle, extracellular matrix modulation and growth, epithelial-mesenchymal transition, and metabolism (Data Set7b). In addition, several cytokines that are associated with kidney injury are upregulated in the early phases of PKD, such as osteopontin (*OPN*) and growth differentiation factor 15 (*GDF15*).

Experimental Validation of the Injury Component in Independent Samples

To validate our results, we generated an independent set of adult-onset slow-progressing inducible *Pkd1*-deletion mice as a PKD model not included in the generation of the PKD Signature. We harvested mice at five different time points (1, 2, 5, and 10 wk after injury and at kidney failure) (Fig. 5*A*).

These samples were analyzed by Fluidigm qPCR chip to quantify mRNA levels of selected genes; inflammatory response (*PcdH7* and *Stat3*), hypoxia (*Akap12* and *Anxa2*), epithelial-mesenchymal transition (*Dpysl3* and *Tnfrsf12a*), TNF-α /NF-κB signaling (*Socs3*), coagulation (*Fgg*), transcription factor (*Glis2*), and transporters (*Cp*). Additionally, we have included an age-matched set of mouse samples, i.e., WT mice with and without treatment with the nephrotoxic compound DCVC, to reflect the injury repair component. Normalized Ct values were tested for statistical significance between the groups (mutant vs. WT, and WT + DCVC vs. WT). All genes from the PKD Signature were significantly dysregulated between mutant and WTs and all genes we classified as injury repair related were significantly dysregulated between (WT and WT + DCVC) (Fig. 5*B*).

Macrophages in PKD

Macrophages have important roles in renal injury repair and PKD (2, 50, 60). Having identified the injury repair genes of the PKD Signature, we proceeded to identify novel mac-

Table 3. *Top 5 annotations for genes in the injury repair PKD Signature component*

Term Name	Count	FDR	Genes (Gene Symbol)
TNFA_SIGNALING_VIA_NFKB	43	1.9E-38	*CD44, GADD45B, RHOB, PLAUR, JUN, CYR61, CXCL1, TNC, MAFF, F3, KLF6, CSF1, MYC, NFIL3, PHLDA1, GCH1, CDKN1A, BTG2, HBEGF, SPHK1, TLR2, ICAM1, CCL2, KYNU, ETS2, PPP1R15A, BTG1, FOSL2, BIRC3, MCL1, PLK2, IER5, SOCS3, CXCL3, BCL3, CEBPB, JUNB, CXCL2, TGIF1, CEBPD, EGR1, PANX1, LITAF*
EPITHELIAL_MESENCHYMAL_TRANSITION	37	1.44E-30	*CD44, GADD45B, RHOB, PLAUR, JUN, CYR61, CXCL1, TNC, TIMP1, SPARC, FN1, FBN1, THBS1, SPP1, ITGA5, DCN, TGFBI, TNFRSF12A, SDC1, COL4A2, COL3A1, COL1A1, ITGB1, ITGB5, TAGLN, TPM2, COL1A2, LAMC2, FSTL1, BASP1, DPYSL3, LOXL2, MGP, PCOLCE, SFRP4, TPM4, VIM*
COAGULATION	25	3.65E-20	*TIMP1, SPARC, FN1, FBN1, THBS1, MAFF, F3, MEP1A, KLF7, ANXA1, CFH, MMP9, DUSP6, CAPN5, PF4, C3, CSRP1, FGA, FGG, MSRB2, CFI, APOC3, P2RY1, PROC, CAPN2*
INFLAMMATORY_RESPONSE	28	8.01E-20	*TIMP1, F3, MEP1A, PLAUR, ITGA5, KLF6, CSF1, MYC, GCH1, CDKN1A, BTG2, HBEGF, SPHK1, TLR2, ICAM1, CCL2, RGS16, TNFRSF1B, IL4R, TPBG, CD14, OSMR, TAPBP, BST2, RHOG, SLC7A1, PDPN, AXL*
XENOBIOTIC_METABOLISM	28	8.01E-20	*GCH1, KYNU, ETS2, GABARAPL1, AHCY, RETSAT, CROT, HSD11B1, ALDH9A1, TNFRSF1A, CAT, ACOX1, POR, APOE, FMO1, BPHL, PDK4, SLC6A12, TMBIM6, ESR1, SLC46A3, GSS, ARG2, MAN1A1, NDRG2, SLC12A4, NFS1, ENTPD5*

Count represents the number of genes that belong to a functional annotation term and the tested PKD Signature component. FDR, false discovery rate, based on MSigDB.

rophage-related molecular pathways involved in PKD progression. We used the data of Clements et al. (11), which contain unique expression profiles of different macrophage populations upon renal injury induction. They identified three distinct macrophage populations: the "CD11b+/Ly6C^{high}" population associated with the onset of renal injury and increase in proinflammatory cytokines, the "CD11b+/Ly6C^{intermediate}" population that peaked during kidney repair, and the "CD11b+/Ly6C^{low}" population that emerged with developing renal fibrosis. We looked for genes that are upregulated in the PKD Signature and uniquely upregulated in each of the three macrophage populations by selecting, for each population, the genes that are upregulated to the other two (logFC ≥ 2, $P < 0.05$). The most enriched populations in PKD are the "CD11b+/Ly6C^{intermediate}" and "CD11b+/Ly6C^{low}" populations, both are two times more enriched in the injury repair genes when compared with the noninjury repair genes (Fig. 4*B*). As these are macrophages activated upon injury induction, stronger enrichment in the injury repair component in contrast to the noninjury repair component is expected and observed.

Using these data sets, we identified several novel genes that are involved in macrophage-related wound healing and fibrosis processes in PKD (Data Set8a). KEGG pathway annotations revealed that these genes are involved in pathways related to extracellular matrix-receptor interaction, focal adhesion, regulation of actin cytoskeleton, and cytokine-cytokine receptor interaction (Data Set8b).

Comparison of the PKD Signature with Other Kidney Diseases

We compared the PKD Signature with other kidney diseases, for which expression profiling studies have been published in the literature (1, 7, 63, 72) (Table 4). The *Fh1* knockout hereditary leiomyomatosis and renal cell cancer (HLRCC) model that develops renal cysts (49), had the highest enrichment with the PKD Signature with an RF of 3.3. In addition, the overlap of other kidney diseases such as glomerulonephritis (63) and diabetic nephropathy (72), with the PKD Signature increases as the severity of these diseases increases, evident by data acquired from 4- and 8-wk-old glomerulonephritis mice (RF of 1.7 and 2.9, respectively) and 8- and 32-wk-old diabetic nephropathy mice (RF of 1.8 and 2.8, respectively). Functional annotation tests of the genes of the more advanced disease stages revealed their involvement in functions related to the immune system and epithelial-mesenchymal transition, suggesting, that most of the overlap of the PKD Signature with other kidney diseases is related to injury repair and inflammation. To test this, we compared the enrichment of each disease with both the injury repair genes and the noninjury repair genes of the PKD Signature (Fig. 4*C*). The results confirmed that the overlap with the injury repair component of PKD was twice as large as the overlap with the noninjury repair component.

Utilizing the large variety of data sets that we have compiled and compared with the PKD Signature, we are able to classify the PKD Signature genes into different categories. These categories are based on the level of commonality of the PKD Signatures genes to other kidney diseases and injury repair processes (Fig. 4*D*). Interestingly, the unique PKD Signature genes are the genes with the least known functional annotations (<12% of genes mapped to annotation terms) (Fig. 4*D*).

DISCUSSION

In this work, we have created a robust PKD Signature that would help in the ongoing translational research efforts to find novel treatments for PKD patients. We followed a meta-analysis approach that combined available PKD expression profiling studies with a new data set that we contributed. Given that the data sets that we could include are limited and variable, we focused on creating a PKD Signature that zooms in on the commonalities of the disease and included genes that are consistently dysregulated across the different studies. Although in this approach we do not guarantee to include all genes involved or dysregulated in PKD, we managed to include highly relevant PKD genes. To corroborate this, we tested the PKD Signature on an independent PKD data set that was not

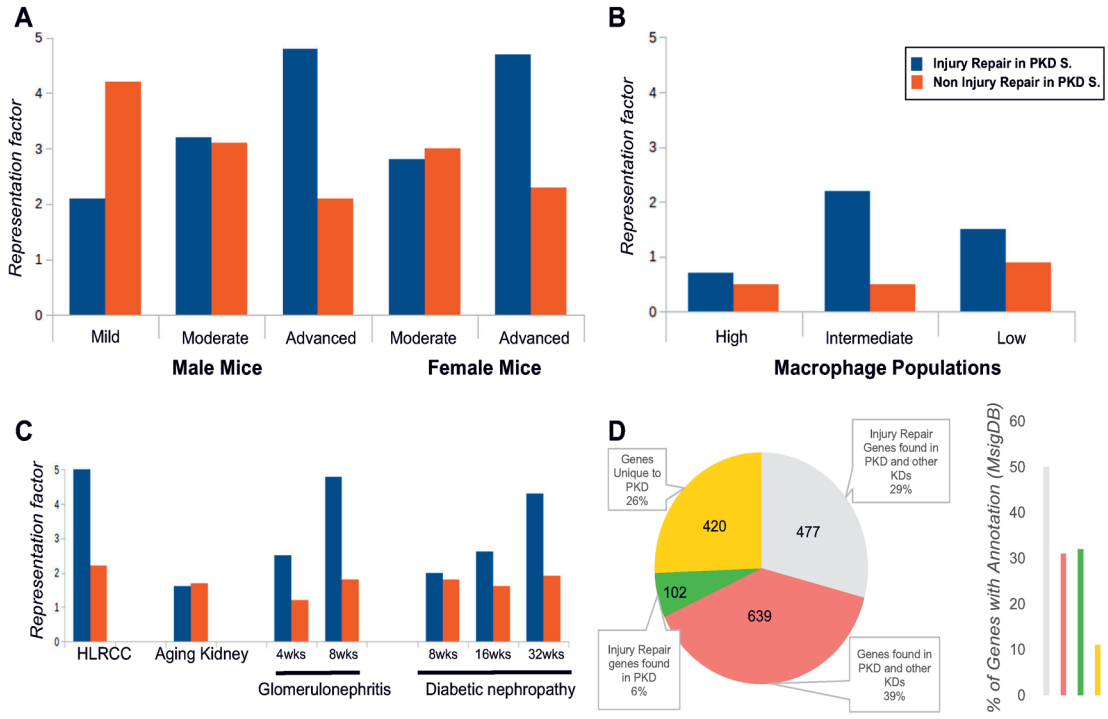

Fig. 4. Validation of injury repair processes in PKD Signature. *A*: enrichment analysis (RF, *y*-axis) of the injury repair (blue) and noninjury repair (red) components of the PKD Signature (PKD S.) in the different phases of the severity of PKD disease in *Pkd1cko* mice (38). *B*: bar chart showing the strong enrichment of different macrophage population [$CD11b^{+}/Ly6C^{high,\ intermediate,\ and\ low}$, Clements et al. (11)] gene signatures activated upon injury induction in mouse models in the injury repair (blue) and noninjury repair (red) components of the PKD Signature. *C*: the enrichment of the injury repair (blue) and noninjury repair (red) components of the PKD Signature in expression profiles from other kidney diseases. *D*: characterization of the PKD Signature into groups based on the overlap with renal injury repair and other kidney diseases. On the *left*, a pie chart showing the total number of genes within each of the 4 groups, and on the *right*, a bar chart reflecting the percentage of genes that matched to at least one annotation term based on MsigDB Hallmarks and Reactome pathway database after running the enrichment comparison with the following settings (FDR < 0.05, top 20 terms).

used in the making of the PKD Signature. This analysis revealed that the PKD Signature genes are three times more enriched in an independent PKD study compared with the PKD genes that were excluded from our PKD Signature. In addition, significant enrichments were observed in mild, moderate, and advanced stages of the disease in both males and females. These results reveal the robustness of the PKD Signature and support the likelihood of preserved key disease mechanisms between genders. We also experimentally confirmed the dysregulation of a selection of genes from the PKD Signature in an independent PKD model using qPCR. Our functional annotation of the PKD Signature revealed the dysregulation of many PKD-linked pathways, such as epithelial-mesenchymal transition (8), TGF-β signaling pathway (8, 17), cell cycle, JAK/STAT signaling pathway (17, 39, 52) and mammalian target of rapamycin (mTOR) signaling [42, 55; in addition to downregulated genes in molecular transport (64) and a large set of genes involved in metabolic pathways (5, 53)].

Being motivated by the rich literature suggesting a strong relationship between PKD and injury repair processes, we went on to characterize the injury repair involved genes in the PKD Signature. We used a novel approach to combine experimental and literature-based renal injury repair profiles. We also contributed a new data set of injury induced mouse model that covers the late-injury repair response phase. Our results are in-line with these suggestions, where we were able to link 35% (581 genes) of the PKD Signature genes to injury repair response. This overlap is two times more enriched than randomly expected (RF = 2). Interestingly, these injury repair genes are involved in more than 65% of the functional annotations attributed to PKD. The most enriched functions of the PKD Signature injury repair genes include epithelial-mesenchymal transition, proliferation, apoptosis, hypoxia, inflammatory response, TNF-α/NF-κB signaling, and glycolysis. Furthermore, we showed that the injury repair genes are enriched in all stages of the PKD disease and their enrichment becomes stronger as the disease progresses toward a more severe state. This is in accordance with our understanding of the disease. As PKD progresses, cysts grow in size, putting pressure on nearby cells and giving rise to new cysts. Oxidative stress increases with disease progression and injury repair processes become significantly evident with visible macrophage infiltration and fibrosis taking place at the cyst site. The functional annotation profiles of the PKD Signature injury repair genes reflect this. In mild and moderate phases of the disease, injury repair genes are associated with cell cycle-related events (i.e., *BIRC5*, *MCM2*, *MCM5*, *PLK1*, and *CDKN1A*), genes involved in extracellular matrix development and morphogenesis (i.e.,

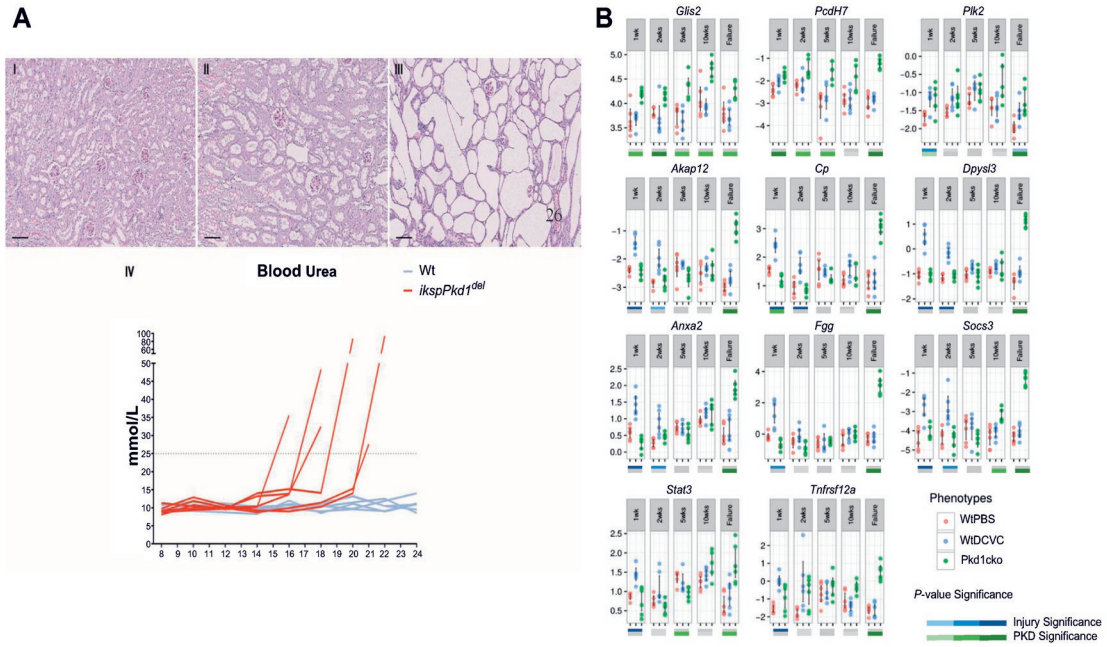

Fig. 5. Validation of the PKD Signature in an adult onset iKsp-*Pkd1*^{del} model and in a nephrotoxic injury model. *A*: Cyst progression in the adult iKsp-*Pkd1*^{del} mice. *I–III*: PAS staining on formalin-fixed, paraffin-embedded kidney sections; scale bar = 100 um. Mild tubular dilation at 5 wk (*I*), 10 wk after gene disruption (*II*), and many cysts at kidney failure (*III*). *IV*: BU levels were used to assess renal failure and are presented for individual mice. A slow progression of the disease was observed, with median duration until kidney failure of 19 wk. *B*: genes selected randomly from the PKD Signature, were subjected to qPCR on the iKsp-*Pkd1*^{del} mice model described in *A* and age-matched WTs at 1, 2, 5, 10 wk after gene knockout and at kidney failure. Normalized Ct values (cycle threshold values) are plotted (log2 scale) for each gene separately across five measurement time points for three types of samples: wild-type mice treated with saline (WTPBS, red), iKsp-*Pkd1*^{del} mice treated with saline (Pkd1cko, green), and wild-type mice treated with DCVC (WTDCVC, blue). For each time point, the significance of dysregulation in RNA-levels with respect to WTPBS mice was tested using Student's *t*-test. Grey bars below the *x*-axis are used in case of no significance ($P \geq 0.01$) and darker shades of colors (green for PKD and blue for injury) are used to denote stronger dysregulation ($P < 0.01$, $P < 0.001$, and $P < 0.0001$ for the 3 different shades of color).

OPN, *TPM4*, *TGFBI*, and *TNFRSF12A*), and genes involved in transport of cations/anions and amino acids/oligopeptides (i.e., *SLC25A10*, *SLC38A2*, and *SLC6A12*) and metabolism (i.e., *CHPF* and *LGALS3*). On the other hand, injury repair genes at the late phases of PKD are involved in the negative regulation of apoptosis, hypoxia, inflammatory response and TNF-α/NF-κB signaling. Additionally, the upregulation of renal injury markers osteopontin (*OPN*) in the moderate phase and kidney injury molecule 1 (*KIM-1* and *HAVCR1*) in the advanced phase of PKD, confirms the involvement renal injury repair processes in disease progression (66).

Utilizing the data of Clements et al. (11) we looked for PKD genes that are involved in macrophage-related wound healing and fibrosis events after injury induction. Our results revealed that macrophage-related genes activated upon injury induction are more enriched in the PKD Signature injury repair genes than the noninjury repair genes. This is especially evident in the macrophage populations CD11b^{+}/Ly6C^{intermediate} and CD11b^{+}/Ly6C^{low}, suggesting a role for PKD injury repair genes in macrophage-related wound healing and fibrosis events. Syndecan-1 (*SDC1*) (36), secreted protein acidic and cysteine rich (*SPARC*) (48), and collagen type i α2 chain (*COL1A2*) (14) are three known fibrosis genes found in the injury repair PKD Signature genes. However, it remains unclear whether these genes are reflecting only the expression in macrophages or also expression in the epithelium. As macrophages are known to contribute to PKD's pathology, further research is needed to determine their clinical significance in PKD treatment.

Table 4. *Published Expression Profiling Studies used in the comparison of the PKD Signature to other diseases*

Authors	Organism	Accession	Disease	Mouse Model
Adam et al. (1)	Mouse	GSE10989	HLRCC	*Fh1* knockout
Braun et al. (7)	Mouse	GSE3219	Aging kidney	Aged wild-type mice
Tamura et al. (63)	Mouse	GSE45005	Glomerulonephritis	ICGN mice
Yang et al. (72)	Mouse	GSE20844	Diabetic nephropathy	OVE26 mice

HLRCC, hereditary leiomyomatosis and renal cell cancer; ICGN mice, ICR-derived glomerulonephritis.

We validated the PKD Signature genes in a second PKD model with more sampling time points, using a Fluidigm qPCR chip. We also included WT and WT + DCVC samples at matching time points. This analysis showed that our classification into "injury repair" and "noninjury repair" groups was predictive. For instance, *Glis2* an important gene in kidney function (22, 26) is part of the PKD Signature and was consistently upregulated in all qPCR measured time points of the iKsp-*Pkd1*^{del}. Our computational analysis did not include it

as part of the injury repair genes in the PKD Signature, and this was further confirmed in the qPCR results as it did not respond to DCVC injury induction in WT mice. *Plk2*, on the other hand, is upregulated in the early PKD time point and in the injury induced mice, confirming our classification of *Plk2* as an injury repair related gene.

In conclusion, our computational methodology combined by our experimental validation in an independent mouse model identified a robust expression signature of PKD. We provide an extensive meta-analysis of PKD transcriptional profile and characterization of the injury repair genes involved. We believe that this study can be used to improve our understanding of the disease and how altered injury repair processes augment it over time. Comparing the PKD Signature to other kidney diseases revealed that the injury repair genes in the PKD Signature account, in large part to the evident overlap between PKD and other kidney diseases, revealing the significance of injury repair genes and pathways in the development and progression of kidney diseases. Novel drug targets can be identified from the profiles affecting the common subset of injury repair processes or affecting the PKD-unique targets. This may allow for drug repurposing between renal diseases and/or the identification of PKD unique agents that modulate the regenerative process in PKD patients and may lead to the much sought after treatment for PKD patients.

ACKNOWLEDGMENTS

We thank Kristina M. Hettne, Leo Price, and Freek van Eeden for critically reading and editing the manuscript.

GRANTS

The research leading to these results has received funding from the People Program (Marie Curie Actions) of the European Union's Seventh Framework Program FP7/2077-2013 under Research Executive Agency Grant Agreement 317246 and the Dutch Technology Foundation Stichting Technische Wetenschappen Project 11823, which is part of The Netherlands Organization for Scientific Research.

DISCLOSURES

No conflicts of interest, financial or otherwise, are declared by the authors.

AUTHOR CONTRIBUTIONS

T.B.M., M.R., D.J.P., and P.A.t.H. conceived and designed research; T.B.M., C.F., W.N.L., P.R., and Z.G. performed experiments; T.B.M., C.F., W.N.L., P.R., Z.G., and M.R. analyzed data; T.B.M., C.F., W.N.L., and D.J.P. interpreted results of experiments; T.B.M. and C.F. prepared figures; T.B.M. drafted manuscript; T.B.M., C.F., W.N.L., Z.G., M.R., D.J.P., and P.A.t.H. edited and revised manuscript; T.B.M., C.F., W.N.L., P.R., Z.G., M.R., D.J.P., and P.A.t.H. approved final version of manuscript.

REFERENCES

1. **Adam J, Hatipoglu E, O'Flaherty L, Ternette N, Sahgal N, Lockstone H, Baban D, Nye E, Stamp GW, Wolhuter K, Stevens M, Fischer R, Carmeliet P, Maxwell PH, Pugh CW, Frizzell N, Soga T, Kessler BM, El-Bahrawy M, Ratcliffe PJ, Pollard PJ.** Renal cyst formation in Fh1-deficient mice is independent of the Hif/Phd pathway: roles for fumarate in KEAP1 succination and Nrf2 signaling. *Cancer Cell* 20: 524–537, 2011. doi:10.1016/j.ccr.2011.09.006.
2. **Anders HJ, Ryu M.** Renal microenvironments and macrophage phenotypes determine progression or resolution of renal inflammation and fibrosis. *Kidney Int* 80: 915–925, 2011. doi:10.1038/ki.2011.217.
3. **Anders S, Pyl PT, Huber W.** HTSeq–a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31: 166–169, 2015. doi:10.1093/bioinformatics/btu638.
4. **The International Polycystic Kidney Disease Consortium.** Polycystic kidney disease: the complete structure of the PKD1 gene and its protein. The International Polycystic Kidney Disease Consortium. *Cell* 81: 289–298, 1995. doi:10.1016/0092-8674(95)90339-9.
5. **Aukema HM, Yamaguchi T, Takahashi H, Celi B, Holub BJ.** Abnormal lipid and fatty acid compositions of kidneys from mice with polycystic kidney disease. *Lipids* 27: 429–435, 1992. doi:10.1007/BF02536384.
6. **Bell PD, Fitzgibbon W, Sas K, Stenbit AE, Amria M, Houston A, Reichert R, Gilley S, Siegal GP, Bissler J, Bilgen M, Chou PC, Guay-Woodford L, Yoder B, Haycraft CJ, Siroky B.** Loss of primary cilia upregulates renal hypertrophic signaling and promotes cystogenesis. *J Am Soc Nephrol* 22: 839–848, 2011. doi:10.1681/ASN.2010050526.
7. **Braun F, Rinschen MM, Bartels V, Frommolt P, Habermann B, Hoeijmakers JH, Schumacher B, Dollé ME, Müller RU, Benzing T, Schermer B, Kurschat CE.** Altered lipid metabolism in the aging kidney identified by three layered omic analysis. *Aging* (Albany NY) 8: 441–457, 2016. doi:10.18632/aging.100900.
8. **Chea SW, Lee KB.** TGF-beta mediated epithelial-mesenchymal transition in autosomal dominant polycystic kidney disease. *Yonsei Med J* 50: 105–111, 2009. doi:10.3349/ymj.2009.50.1.105.
9. **Chen MJ, Wong CH, Peng ZF, Manikandan J, Melendez AJ, Tan TM, Crack PJ, Cheung NS.** A global transcriptomic view of the multifaceted role of glutathione peroxidase-1 in cerebral ischemic-reperfusion injury. *Free Radic Biol Med* 50: 736–748, 2011. doi:10.1016/j.freeradbiomed.2010.12.025.
10. **Chen WC, Tzeng YS, Li H.** Gene expression in early and progression phases of autosomal dominant polycystic kidney disease. *BMC Res Notes* 1: 131, 2008. doi:10.1186/1756-0500-1-131.
11. **Clements M, Gershenovich M, Chaber C, Campos-Rivera J, Du P, Zhang M, Ledbetter S, Zuk A.** Differential Ly6C expression after renal ischemia-reperfusion identifies unique macrophage populations. *J Am Soc Nephrol* 27: 159–170, 2016. doi:10.1681/ASN.2014111138.
12. **Correa-Costa M, Azevedo H, Amano MT, Gonçalves GM, Hyane MI, Cenedeze MA, Renesto PG, Pacheco-Silva A, Moreira-Filho CA, Câmara NO.** Transcriptome analysis of renal ischemia/reperfusion injury and its modulation by ischemic pre-conditioning or hemin treatment. *PLoS One* 7: e49569, 2012. doi:10.1371/journal.pone.0049569.
13. **Ferreira FM, Watanabe EH, Onuchic LF.** Polycystins and molecular basis of autosomal dominant polycystic kidney disease. In: *Polycystic Kidney Disease*, edited by Li X. Brisbane, Australia: Codon, 2015.
14. **Fragiadaki M, Witherden AS, Kaneko T, Sonnylal S, Pusey CD, Bou-Gharios G, Mason RM.** Interstitial fibrosis is associated with increased COL1A2 transcription in AA-injured renal tubular epithelial cells in vivo. *Matrix Biol* 30: 396–403, 2011. doi:10.1016/j.matbio.2011.07.004.
15. **Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J.** Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* 5: R80, 2004. doi:10.1186/gb-2004-5-10-r80.
16. **Happé H, Leonhard WN, van der Wal A, van de Water B, Lantinga-van Leeuwen IS, Breuning MH, de Heer E, Peters DJ.** Toxic tubular injury in kidneys from Pkd1-deletion mice accelerates cystogenesis accompanied by dysregulated planar cell polarity and canonical Wnt signaling pathways. *Hum Mol Genet* 18: 2532–2542, 2009. doi:10.1093/hmg/ddp190.
17. **Hassane S, Leonhard WN, van der Wal A, Hawinkels LJ, Lantinga-van Leeuwen IS, ten Dijke P, Breuning MH, de Heer E, Peters DJ.** Elevated TGFbeta-Smad signalling in experimental Pkd1 models and human patients with polycystic kidney disease. *J Pathol* 222: 21–31, 2010. doi:10.1002/path.2734.

17a. **Hettne KM, van Schouwen R, Mina E, van der Horst E, Thompson M, Rajaram Kaliyaperumal R, Mons B, van Mulligen E, Kors JA, Roos M.** Explain your data by Concept Profile Analysis Web Services [version 1; referees: 2 approved with reservations]. *F1000 Res* 3: 173, 2014. doi:10.12688/f1000research.4830.1.

18. **Huang W, Sherman BT, Lempicki RA.** Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* 37: 1–13, 2009. doi:10.1093/nar/gkn923.
19. **Huang W, Sherman BT, Lempicki RA.** Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 4: 44–57, 2009. doi:10.1038/nprot.2008.211.
20. **Husson H, Manavalan P, Akmaev VR, Russo RJ, Cook B, Richards B, Barberio D, Liu D, Cao X, Landes GM, Wang CJ, Roberts BL, Klinger KW, Grubman SA, Jefferson DM, Ibraghimov-Beskrovnaya**

O. New insights into ADPKD molecular pathways using combination of SAGE and microarray technologies. *Genomics* 84: 497–510, 2004. doi:10.1016/j.ygeno.2004.03.009.

21. **Jelier R, Schuemie MJ, Veldhoven A, Dorssers LC, Jenster G, Kors JA.** Anni 2.0: a multipurpose text-mining tool for the life sciences. *Genome Biol* 9: R96, 2008. doi:10.1186/gb-2008-9-6-r96.
22. **Kang HS, ZeRuth G, Lichti-Kaiser K, Vasanth S, Yin Z, Kim YS, Jetten AM.** Gli-similar (Glis) Krüppel-like zinc finger proteins: insights into their physiological functions and critical roles in neonatal diabetes and cystic renal disease. *Histol Histopathol* 25: 1481–1496, 2010. doi:10.14670/HH-25.1481.
23. **Kauffmann A, Rayner TF, Parkinson H, Kapushesky M, Lukk M, Brazma A, Huber W.** Importing ArrayExpress data sets into R/Bioconductor. *Bioinformatics* 25: 2092–2094, 2009. doi:10.1093/bioinformatics/btp354. https://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=19505942&dopt=Abstract
24. **Kennefick TM, Al-Nimri MA, Oyama TT, Thompson MM, Kelly FJ, Chapman JG, Anderson S.** Hypertension and renal injury in experimental polycystic kidney disease. *Kidney Int* 56: 2181–2190, 1999. doi:10.1046/j.1523-1755.1999.00783.x.
25. **Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL.** TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* 14: R36, 2013. doi:10.1186/gb-2013-14-4-r36.
26. **Kim YS, Kang HS, Herbert R, Beak JY, Collins JB, Grissom SF, Jetten AM.** Kruppel-like zinc finger protein Glis2 is essential for the maintenance of normal renal functions. *Mol Cell Biol* 28: 2358–2367, 2008. doi:10.1128/MCB.01722-07.
27. **Kolesnikov N, Hastings E, Keays M, Melnichuk O, Tang YA, Williams E, Dylag M, Kurbatova N, Brandizi M, Burdett T, Megy K, Pilicheva E, Rustici G, Tikhonov A, Parkinson H, Petryszak R, Sarkans U, Brazma A.** ArrayExpress update–simplifying data submissions. *Nucleic Acids Res* 43: D1113–D1116, 2015. doi:10.1093/nar/gku1057.
28. **Krishnamoorthy A, Ajay AK, Hoffmann D, Kim TM, Ramirez V, Campanholle G, Bobadilla NA, Waikar SS, Vaidya VS.** Fibrinogen β-derived Bβ(15–42) peptide protects against kidney ischemia/reperfusion injury. *Blood* 118: 1934–1942, 2011. doi:10.1182/blood-2011-02-338061.
30. **Kugita M, Nishii K, Morita M, Yoshihara D, Kowa-Sugiyama H, Yamada K, Yamaguchi T, Wallace DP, Calvet JP, Kurahashi H, Nagao S.** Global gene expression profiling in early-stage polycystic kidney disease in the Han:SPRD Cy rat identifies a role for RXR signaling. *Am J Physiol Renal Physiol* 300: F177–F188, 2011. doi:10.1152/ajprenal.00470.2010.
31. **Langmead B, Salzberg SL.** Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9: 357–359, 2012. doi:10.1038/nmeth.1923.
32. **Lantinga-van Leeuwen IS, Leonhard WN, van der Wal A, Breuning MH, de Heer E, Peters DJ.** Kidney-specific inactivation of the Pkd1 gene induces rapid cyst formation in developing kidneys and a slow onset of disease in adult mice. *Hum Mol Genet* 16: 3188–3196, 2007. doi:10.1093/hmg/ddm299.
33. **Leonhard WN, Zandbergen M, Veraar K, van den Berg S, van der Weerd L, Breuning M, de Heer E, Peters DJ.** Scattered deletion of PKD1 in kidneys causes a cystic snowball effect and recapitulates polycystic kidney disease. *J Am Soc Nephrol* 26: 1322–1333, 2015. doi:10.1681/ASN.2013080864.
34. **Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P.** The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst* 1: 417–425, 2015. doi:10.1016/j.cels.2015.12.004.
35. **Liu J, Krautzberger AM, Sui SH, Hofmann OM, Chen Y, Baetscher M, Grgic I, Kumar S, Humphreys BD, Hide WA, McMahon AP.** Cell-specific translational profiling in acute kidney injury. *J Clin Invest* 124: 1242–1254, 2014. doi:10.1172/JCI72126.
36. **Masola V, Gambaro G, Tibaldi E, Brunati AM, Gastaldello A, D'Angelo A, Onisto M, Lupo A.** Heparanase and syndecan-1 interplay orchestrates fibroblast growth factor-2-induced epithelial-mesenchymal transition in renal tubular cells. *J Biol Chem* 287: 1478–1488, 2012. doi:10.1074/jbc.M111.279836.
37. **Menezes LF, Germino GG.** Systems biology of polycystic kidney disease: a critical review. *Wiley Interdiscip Rev Syst Biol Med* 7: 39–52, 2015. doi:10.1002/wsbm.1289.
38. **Menezes LF, Lin CC, Zhou F, Germino GG.** Fatty acid oxidation is impaired in an orthologous mouse model of autosomal dominant polycystic kidney disease. *EBioMedicine* 5: 183–192, 2016. doi:10.1016/j.ebiom.2016.01.027.
39. **Menezes LF, Zhou F, Patterson AD, Piontek KB, Krausz KW, Gonzalez FJ, Germino GG.** Network analysis of a Pkd1-mouse model of autosomal dominant polycystic kidney disease identifies HNF4α as a disease modifier. *PLoS Genet* 8: e1003053, 2012. doi:10.1371/journal.pgen.1003053.
40. **Mochizuki T, Wu G, Hayashi T, Xenophontos SL, Veldhuisen B, Saris JJ, Reynolds DM, Cai Y, Gabow PA, Pierides A, Kimberling WJ, Breuning MH, Deltas CC, Peters DJ, Somlo S.** PKD2, a gene for polycystic kidney disease that encodes an integral membrane protein. *Science* 272: 1339–1342, 1996. doi:10.1126/science.272.5266.1339.
41. **Mudunuri U, Che A, Yi M, Stephens RM.** bioDBnet: the biological database network. *Bioinformatics* 25: 555–556, 2009. doi:10.1093/bioinformatics/btn654.
42. **Novalic Z, van der Wal AM, Leonhard WN, Koehl G, Breuning MH, Geissler EK, de Heer E, Peters DJ.** Dose-dependent effects of sirolimus on mTOR signaling and polycystic kidney disease. *J Am Soc Nephrol* 23: 842–853, 2012. doi:10.1681/ASN.2011040340.
43. **O'Meara CC, Hoffman M, Sweeney WE Jr, Tsaih SW, Xiao B, Jacob HJ, Avner ED, Moreno C.** Role of genetic modifiers in an orthologous rat model of ARPKD. *Physiol Genomics* 44: 741–753, 2012. doi:10.1152/physiolgenomics.00187.2011.
44. **Pandey P, Qin S, Ho J, Zhou J, Kreidberg JA.** Systems biology approach to identify transcriptome reprogramming and candidate microRNA targets during the progression of polycystic kidney disease. *BMC Syst Biol* 5: 56, 2011. doi:10.1186/1752-0509-5-56.
45. **Park EY, Seo MJ, Park JH.** Effects of specific genes activating RAGE on polycystic kidney disease. *Am J Nephrol* 32: 169–178, 2010. doi:10.1159/000315859.
46. **Park JY, Schutzer WE, Lindsley JN, Bagby SP, Oyama TT, Anderson S, Weiss RH.** p21 is decreased in polycystic kidney disease and leads to increased epithelial cell cycle progression: roscovitine augments p21 levels. *BMC Nephrol* 8: 12, 2007. doi:10.1186/1471-2369-8-12.
47. **Patel V, Li L, Cobo-Stark P, Shao X, Somlo S, Lin F, Igarashi P.** Acute kidney injury and aberrant planar cell polarity induce cyst formation in mice lacking renal cilia. *Hum Mol Genet* 17: 1578–1590, 2008. doi:10.1093/hmg/ddn045.
48. **Pichler RH, Hugo C, Shankland SJ, Reed MJ, Bassuk JA, Andoh TF, Lombardi DM, Schwartz SM, Bennett WM, Alpers CE, Sage EH, Johnson RJ, Couser WG.** SPARC is expressed in renal interstitial fibrosis and in renal vascular injury. *Kidney Int* 50: 1978–1989, 1996. doi:10.1038/ki.1996.520.
49. **Pollard PJ, Spencer-Dene B, Shukla D, Howarth K, Nye E, El-Bahrawy M, Deheragoda M, Joannou M, McDonald S, Martin A, Igarashi P, Varsani-Brown S, Rosewell I, Poulsom R, Maxwell P, Stamp GW, Tomlinson IP.** Targeted inactivation of fh1 causes proliferative renal cyst development and activation of the hypoxia pathway. *Cancer Cell* 11: 311–319, 2007. doi:10.1016/j.ccr.2007.02.005.
50. **Ricardo SD, van Goor H, Eddy AA.** Macrophage diversity in renal injury and repair. *J Clin Invest* 118: 3522–3530, 2008. doi:10.1172/JCI36150.
51. **Riera M, Burtey S, Fontés M.** Transcriptome analysis of a rat PKD model: Importance of genes involved in extracellular matrix metabolism. *Kidney Int* 69: 1558–1563, 2006. doi:10.1038/sj.ki.5000309.
52. **Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK.** limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 43: e47, 2015. doi:10.1093/nar/gkv007.
53. **Rowe I, Chiaravalli M, Mannella V, Ulisse V, Quilici G, Pema M, Song XW, Xu H, Mari S, Qian F, Pei Y, Musco G, Boletta A.** Defective glucose metabolism in polycystic kidney disease identifies a new therapeutic strategy. *Nat Med* 19: 488–493, 2013. doi:10.1038/nm.3092.
54. **Schieren G, Rumberger B, Klein M, Kreutz C, Wilpert J, Geyer M, Faller D, Timmer J, Quack I, Rump LC, Walz G, Donauer J.** Gene profiling of polycystic kidneys. *Nephrol Dial Transplant* 21: 1816–1824, 2006. doi:10.1093/ndt/gfl071.
55. **Shillingford JM, Murcia NS, Larson CH, Low SH, Hedgepeth R, Brown N, Flask CA, Novick AC, Goldfarb DA, Kramer-Zucker A, Walz G, Piontek KB, Germino GG, Weimbs T.** The mTOR pathway is regulated by polycystin-1, and its inhibition reverses renal cystogenesis in polycystic kidney disease. *Proc Natl Acad Sci USA* 103: 5466–5471, 2006. doi:10.1073/pnas.0509694103.

56. **Song X, Di Giovanni V, He N, Wang K, Ingram A, Rosenblum ND, Pei Y.** Systems biology of autosomal dominant polycystic kidney disease (ADPKD): computational identification of gene expression pathways and integrated regulatory networks. *Hum Mol Genet* 18: 2328–2343, 2009. doi:10.1093/hmg/ddp165.
57. **Speir RW, Stallings JD, Andrews JM, Gelnett MS, Brand TC, Salgar SK.** Effects of valproic acid and dexamethasone administration on early bio-markers and gene expression profile in acute kidney ischemia-reperfusion injury in the rat. *PLoS One* 10: e0126622, 2015. doi:10.1371/journal.pone.0126622.
58. **Subramanian A, Kuehn H, Gould J, Tamayo P, Mesirov JP.** GSEA-P: a desktop application for Gene Set Enrichment Analysis. *Bioinformatics* 23: 3251–3253, 2007. doi:10.1093/bioinformatics/btm369.
59. **Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP.** Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 102: 15545–15550, 2005. doi:10.1073/pnas.0506580102.
60. **Swenson-Fields KI, Vivian CJ, Salah SM, Peda JD, Davis BM, van Rooijen N, Wallace DP, Fields TA.** Macrophages promote polycystic kidney disease progression. *Kidney Int* 83: 855–864, 2013. doi:10.1038/ki.2012.446.
61. **Ta MH, Harris DC, Rangan GK.** Role of interstitial inflammation in the pathogenesis of polycystic kidney disease. *Nephrology (*Carlton) 18: 317–330, 2013. doi:10.1111/nep.12045.
62. **Takakura A, Contrino L, Zhou X, Bonventre JV, Sun Y, Humphreys BD, Zhou J.** Renal injury is a third hit promoting rapid development of adult polycystic kidney disease. *Hum Mol Genet* 18: 2523–2531, 2009. doi:10.1093/hmg/ddp147.
63. **Tamura K, Uchio-Yamada K, Manabe N, Noto T, Hirota R, Unami A, Matsumoto M, Miyamae Y.** Gene expression analysis detected a low expression level of C1s gene in ICR-derived glomerulonephritis (ICGN) mice. *Nephron, Exp Nephrol* 123: 34–45, 2013. doi:10.1159/000354057.
64. **Terryn S, Ho A, Beauwens R, Devuyst O.** Fluid transport and cystogenesis in autosomal dominant polycystic kidney disease. *Biochim Biophys Acta* 1812: 1314–1321, 2011. doi:10.1016/j.bbadis.2011.01.011.
65. **Torres VE, Chapman AB, Devuyst O, Gansevoort RT, Grantham JJ, Higashihara E, Perrone RD, Krasa HB, Ouyang J, Czerwiec FS; TEMPO 3:4 Trial Investigators.** Tolvaptan in patients with autosomal dominant polycystic kidney disease. *N Engl J Med* 367: 2407–2418, 2012. doi:10.1056/NEJMoa1205511.
66. **Vaidya VS, Ferguson MA, Bonventre JV.** Biomarkers of acute kidney injury. *Annu Rev Pharmacol Toxicol* 48: 463–493, 2008. doi:10.1146/annurev.pharmtox.48.113006.094615.
67. **Ward CJ, Hogan MC, Rossetti S, Walker D, Sneddon T, Wang X, Kubly V, Cunningham JM, Bacallao R, Ishibashi M, Milliner DS, Torres VE, Harris PC.** The gene mutated in autosomal recessive polycystic kidney disease encodes a large, receptor-like protein. *Nat Genet* 30: 259–269, 2002. doi:10.1038/ng833.
68. **Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, Antonarakis SE, Attwood J, Baertsch R, Bailey J, Barlow K, Beck S, Berry E, Birren B, Bloom T, Bork P, Botcherby M, Bray N, Brent MR, Brown DG, Brown SD, Bult C, Burton J, Butler J, Campbell RD, Carninci P, Cawley S, Chiaromonte F, Chinwalla AT, Church DM, Clamp M, Clee C, Collins FS, Cook LL, Copley RR, Coulson A, Couronne O, Cuff J, Curwen V, Cutts T, Daly M, David R, Davies J, Delehaunty KD, Deri J, Dermitzakis ET, Dewey C, Dickens NJ, Diekhans M, Dodge S, Dubchak I, Dunn DM, Eddy SR, Elnitski L, Emes RD, Eswara P, Eyras E, Felsenfeld A, Fewell GA, Flicek P, Foley K, Frankel WN, Fulton LA, Fulton RS, Furey TS, Gage D, Gibbs RA, Glusman G, Gnerre S, Goldman N, Goodstadt L, Grafham D, Graves TA, Green ED, Gregory S, Guigó R, Guyer M, Hardison RC, Haussler D, Hayashizaki Y, Hillier LW, Hinrichs A, Hlavina W, Holzer T, Hsu F, Hua A, Hubbard T, Hunt A, Jackson I, Jaffe DB, Johnson LS, Jones M, Jones TA, Joy A, Kamal M, Karlsson EK, Karolchik D, Kasprzyk A, Kawai J, Keibler E, Kells C, Kent WJ, Kirby A, Kolbe DL, Korf I, Kucherlapati RS, Kulbokas EJ, Kulp D, Landers T, Leger JP, Leonard S, Letunic I, Levine R, Li J, Li M, Lloyd C, Lucas S, Ma B, Maglott DR, Mardis ER, Matthews L, Mauceli E, Mayer JH, McCarthy M, McCombie WR, McLaren S, McLay K, McPherson JD, Meldrim J, Meredith B, Mesirov JP, Miller W, Miner TL, Mongin E, Montgomery KT, Morgan M, Mott R, Mullikin JC, Muzny DM, Nash WE, Nelson JO, Nhan MN, Nicol R, Ning Z, Nusbaum C, O'Connor MJ, Okazaki Y, Oliver K, Overton-Larty E, Pachter L, Parra G, Pepin KH, Peterson J, Pevzner P, Plumb R, Pohl CS, Poliakov A, Ponce TC, Ponting CP, Potter S, Quail M, Reymond A, Roe BA, Roskin KM, Rubin EM, Rust AG, Santos R, Sapojnikov V, Schultz B, Schultz J, Schwartz MS, Schwartz S, Scott C, Seaman S, Searle S, Sharpe T, Sheridan A, Shownkeen R, Sims S, Singer JB, Slater G, Smit A, Smith DR, Spencer B, Stabenau A, Stange-Thomann N, Sugnet C, Suyama M, Tesler G, Thompson J, Torrents D, Trevaskis E, Tromp J, Ucla C, Ureta-Vidal A, Vinson JP, Von Niederhausern AC, Wade CM, Wall M, Weber RJ, Weiss RB, Wendl MC, West AP, Wetterstrand K, Wheeler R, Whelan S, Wierzbowski J, Willey D, Williams S, Wilson RK, Winter E, Worley KC, Wyman D, Yang S, Yang SP, Zdobnov EM, Zody MC, Lander ES; Mouse Genome Sequencing Consortium.** Initial sequencing and comparative analysis of the mouse genome. *Nature* 420: 520–562, 2002. doi:10.1038/nature01262.
70. **Weimbs T.** Polycystic kidney disease and renal injury repair: common pathways, fluid flow, and the function of polycystin-1. *Am J Physiol Renal Physiol* 293: F1423–F1432, 2007. doi:10.1152/ajprenal.00275.2007.
71. **Wilson PD.** Aberrant epithelial cell growth in autosomal dominant polycystic kidney disease. *Am J Kidney Dis* 17: 634–637, 1991. doi:10.1016/S0272-6386(12)80338-6.
72. **Yang L, Brozovic S, Xu J, Long Y, Kralik PM, Waigel S, Zacharias W, Zheng S, Epstein PN.** Inflammatory gene expression in OVE26 diabetic kidney during the development of nephropathy. *Nephron, Exp Nephrol* 119: e8–e20, 2011. doi:10.1159/000324407.
73. **Yuen PS, Jo SK, Holly MK, Hu X, Star RA.** Ischemic and nephrotoxic acute renal failure are distinguished by their broad transcriptomic responses. *Physiol Genomics* 25: 375–386, 2006. doi:10.1152/physiolgenomics.00223.2005.

CHAPTER 3
CHARACTERISATION OF TRANSCRIPTION FACTOR PROFILES IN POLYCYSTIC KIDNEY DISEASE (PKD): IDENTIFICATION AND VALIDATION OF STAT3 AND RUNX1 IN THE INJURY/REPAIR RESPONSE AND PKD PROGRESSION INVOLVEMENT OF INJURY REPAIR PROCESSES

Chiara Formica*, Tareq B. Malas*, Judit Balog, Lotte Verburg, Peter A. C. ‘t Hoen & Dorien J. M. Peters

J Mol Med 97, 1643–1656 (2019). https://doi.org/10.1007/s00109-019-01852-3

November 2019

* Contributed equally

Characterisation of transcription factor profiles in polycystic kidney disease (PKD): identification and validation of STAT3 and RUNX1 in the injury/repair response and PKD progression

Chiara Formica[1] · Tareq Malas[1] · Judit Balog[1] · Lotte Verburg[2] · Peter A. C. 't Hoen[1,3] · Dorien J. M. Peters[1]

Received: 16 May 2019 / Revised: 1 November 2019 / Accepted: 7 November 2019 / Published online: 26 November 2019
© The Author(s) 2019

Abstract

Autosomal dominant polycystic kidney disease (ADPKD) is the most common genetic renal disease, caused in the majority of the cases by a mutation in either the *PKD1* or the *PKD2* gene. ADPKD is characterised by a progressive increase in the number and size of cysts, together with fibrosis and distortion of the renal architecture, over the years. This is accompanied by alterations in a complex network of signalling pathways. However, the underlying molecular mechanisms are not well characterised. Previously, we defined the PKD Signature, a set of genes typically dysregulated in PKD across different disease models from a meta-analysis of expression profiles. Given the importance of transcription factors (TFs) in modulating disease, we focused in this paper on characterising TFs from the PKD Signature. Our results revealed that out of the 1515 genes in the PKD Signature, 92 were TFs with altered expression in PKD, and 32 of those were also implicated in tissue injury/repair mechanisms. Validating the dysregulation of these TFs by qPCR in independent PKD and injury models largely confirmed these findings. STAT3 and RUNX1 displayed the strongest activation in cystic kidneys, as demonstrated by chromatin immunoprecipitation (ChIP) followed by qPCR. Using immunohistochemistry, we showed a dramatic increase of expression after renal injury in mice and cystic renal tissue of mice and humans. Our results suggest a role for STAT3 and RUNX1 and their downstream targets in the aetiology of ADPKD and indicate that the meta-analysis approach is a viable strategy for new target discovery in PKD.

Key messages

- We identified a list of transcription factors (TFs) commonly dysregulated in ADPKD.
- Out of the 92 TFs identified in the PKD Signature, 35% are also involved in injury/repair processes.
- STAT3 and RUNX1 are the most significantly dysregulated TFs after injury and during PKD progression.
- STAT3 and RUNX1 activity is increased in cystic compared to non-cystic mouse kidneys.
- Increased expression of STAT3 and RUNX1 is observed in the nuclei of renal epithelial cells, also in human ADPKD samples.

Keywords Autosomal dominant polycystic kidney disease · kidney injury · Gene expression · Transcription factors · Chromatin immunoprecipitation

Chiara Formica and Tareq Malas contributed equally to this work.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00109-019-01852-3) contains supplementary material, which is available to authorized users.

✉ Dorien J. M. Peters
d.j.m.peters@lumc.nl

[1] Department of Human Genetics, Leiden University Medical Center, Einthovenweg 20, 2333, ZC Leiden, The Netherlands

[2] Department of Pathology, Leiden University Medical Center, Albinusdreef 2, 2333, ZA Leiden, The Netherlands

[3] Centre for Molecular and Biomolecular Informatics, Radboud Institute for Molecular Life Sciences, Radboud University Medical Center Nijmegen, Geert Grooteplein Zuid 26/28, 6525, GA Nijmegen, The Netherlands

Introduction

Autosomal dominant polycystic kidney disease (ADPKD) is a genetic disease characterised by the formation of fluid-filled renal cysts. Cyst formation and cyst growth are accompanied by inflammation and fibrosis, leading to kidney failure. In the majority of cases, ADPKD is caused by a mutation in the *PKD1* gene or, less frequently, in the *PKD2* gene. Nevertheless, ADPKD is a complex disease which involves the dysregulation of many different signalling pathways [1], and the molecular mechanisms involved in disease progression are not entirely understood. Currently, the vasopressin V2-receptor antagonist, tolvaptan, is the only approved treatment in Europe but only for selected patients. More generic and definitive treatment is still missing.

Both environmental and genetic factors can be considered disease modifiers in ADPKD [1, 2]. An important one is renal injury, shown to accelerate cyst formation and expansion in different mouse models [3, 4]. Recently, we showed that renal injury shares molecular processes with ADPKD progression. Using a meta-analysis approach, we identified a set of genes dysregulated in a variety of PKD models during disease progression, which we called the "PKD Signature". About 35% of these genes were found to be also implicated in injury/repair mechanisms, confirming the strong relation between ADPKD and injury [5].

Transcription factor (TF) proteins are master regulators of transcription, which control the expression of genes involved in the establishment and maintenance of cell states, in physiological and pathological situations. Dysregulation of TFs levels and/or activity can lead to the development of a broad range of diseases. Thus, identification of a TFs profile in ADPKD could help to better understand the molecular mechanisms contributing to cyst formation. For this reason, in this study, we focus on the signature of TFs. We identified new PKD-related TFs, and we validated altered expression during ADPKD progression and injury/repair in different mouse models. For two of the identified TFs, STAT3 and RUNX1, we also showed increased activity in mouse cystic kidneys, as well as altered expression in human ADPKD kidneys.

Materials and methods

Identification of transcription factors in PKD

Identification of the PKD Signature was described previously [5]. Briefly, in the previous work, we performed a meta-analysis of PKD expression profiles across different disease models and identified 1515 genes that showed consistent dysregulation across the different PKD studies. We further identified genes involved in injury/repair processes from the PKD Signature by firstly producing injury repair gene profile based on several injury-induced animal models and secondly intersecting the identified PKD Signature and injury repair profiles for the identification of overlapping genes.

In this publication, we used MSigDB's collection of TFs based on Messina et al. [6] and Moreland et al. [7] for the identification of TFs involved in PKD. Furthermore, we identified the transcription factors that are involved in the injury/repair processes of PKD based on the previously identified injury repair profile [5].

The enrichment of TF targets in the PKD Signature was based on the target collections in the ChEA 2016 database [8] that includes TF targets based on experimental evidence. We calculated the enrichment using the representation factor method described below. TFs are considered enriched if they had a representation factor above 1. The representation factor is the number of overlapping genes divided by the expected number of overlapping genes drawn from two independent groups. A representation factor > 1 indicates more overlap than expected of two independent groups, and a representation factor < 1 indicates less overlap than expected. The formula used to calculate the representation factor is $x/(n \times D)/N$, where x = # of genes in common between two groups; n = # of genes in group 1 (the total number of targets calculated per transcription factor based on ChEA 2016 database); D = # of genes in group 2 (the total number of genes in the PKD Signature up (775) or down (740) regulated lists independently); N = total genes, in this case, the 10,271 genes with Entrez IDs.

In silico functional annotation of gene lists

GeneTrail2 v1.6 [9] was used to identify the enriched/significant pathways/functions of the identified gene lists. For all analyses, we used Wikipathways as the primary source of annotation. GeneTrail2 v1.6 was run with the following parameters: overrepresentation analysis (enrichment algorithm); FDR adjustment (adjustment method); significance level at 0.05; and minimum and maximum size of the category equal to 2 and 700, respectively.

Gene expression and statistical analysis of the significance of results

Snap-frozen mouse kidneys were homogenised using MagNa Lyser technology (Roche). Total RNA was isolated using TRI Reagent (Sigma-Aldrich). cDNA synthesis was performed using Transcriptor First Strand cDNA Synthesis Kit (Roche), and qPCR was done using 2× FastStart SYBR Green Master (Roche) according to the manufacturer's protocol. Alternatively, it was performed at GenomeScan (GenomeScan B.V.) using the 96.96 BioMark™ Dynamic Array for real-time PCR (Fluidigm Corporation), as previously described [5]. Gene expression was normalised to the

geometric mean of three housekeeping genes (*Rplp0, Hnrnpa2b1, Ywhaz*) for Fluidigm data and *Hprt* for SYBR-Green data. The output of the Fluidigm assay was normalised and converted into Ct values (cycle threshold). For each transcription factor, a two-way ANOVA was conducted to compare the genotype (PKD vs WT) and the treatment (PBS vs DCVC) effects for each age-matched time points. The computation was made using the *Limma package* [10] in R. A list of primer sequences and TaqMan assays can be found in Supplementary Table 3.

Identification of transcription factors binding sites and primer design

For the TFs that were selected for our ChIP analysis, we identified the binding sites of each TF and its targets by screening the Cistrome database [11] and accessing all studies that performed ChIP-Seq experiments on our selected TFs. We looked for peaks that appeared with an intensity of 10 or higher in more than one ChIP-Seq study. We mapped the *Mus musculus* mm10 genome to the peaks identified using *Peak2Gene* tool that is part of the Cistrome Galaxy tools to identify genes that are within 10,000 base pairs of both ends of the peak. The peaks that did not map to a gene target that is part of the PKD Signature were eliminated. Finally, sorting on the intensity level of the peak, we visualised the top peaks on the UCSC Genome Browser [12] and selected the peaks that had sufficient height over noise levels for qPCR enrichment. We designed primers spanning the TFs binding sites on their putative target genes. The binding sites were generally overlapping with the promoter region of the target genes. As a negative control, we designed primers binding at about 5 kb from the promoter regions where we did not expect to find any TF-binding activity. A list of primers can be found in Supplementary Table 3. Two-way ANOVA with Tukey's multiple comparisons test was performed comparing the input-normalised binding-enrichment of the TFs or the control IgG at the binding site and at the nonbinding sites.

Animal model

All the animal experiments were evaluated and approved by the local animal experimental committee of the Leiden University Medical Center (LUMC) and the Commission Biotechnology in Animals of the Dutch Ministry of Agriculture. Kidney-specific tamoxifen-inducible *Pkd1*-deletion mouse model (iKsp*Pkd1*^{del}) have been described previously [13]. We only used male mice, to reduce variability in disease progression as female mice tend to have a slower and milder progression of the disease compared to male mice [14]. Wt mice have only the LoxP sites around exons 2–11 of the *Pkd1* gene but not the Cre recombinase (*Pkd1*^{loxlox}). For three consecutive days, 5 mg/kg of tamoxifen was administered via oral gavage when mice were 13–14 weeks old. Inactivation of the *Pkd1* gene at this age leads to cyst formation in all the renal tubule segments. A week later, mice were injected intraperitoneally with 15 mg/kg of the nephrotoxic compound S-(1,2-dichlorovinyl)-L-cysteine (DCVC) or vehicle (PBS) as a control. Kidney function was evaluated using blood urea nitrogen (BUN) level as previously described [4]. Renal failure is defined by BUN equal or higher than 25 mmol/l. Mice were sacrificed at 1, 2, 5 and 10 weeks after DCVC and kidney failure. The experimental pipeline has been presented in Formica et al. [15]. The Wt + PBS, Wt + DCVC and *Pkd1* KO + PBS groups have also been used in Malas et al. [5]. At the sacrifice, kidneys were collected and weighed. For RNA and chromatin extraction, kidneys were snap frozen in liquid nitrogen. For immunohistochemistry (IHC) staining, kidneys were preserved in phosphate-buffered 4% formaldehyde solution. A t-test was conducted to compare median survival in PBS-treated versus DCVC-treated mice and BUN in Wt versus iKsp*Pkd1*^{del} mice.

ChIP

Chromatin was isolated from mouse inner-medullary collecting duct (mIMCD3; ATCC, Rockville, USA) cells (about 5×10^6/ml). Briefly, cells were crosslinked with 1% formaldehyde for 10 min at RT then lysed with buffer with protease and phosphatase inhibitors (Roche) as described on Nature protocols (ChIP buffer) [16].

For kidneys' chromatin extraction, snap-frozen kidneys, harvested at end-stage renal disease (ESRD) from Wt mice and iKsp*Pkd1*^{del} mice treated with DCVC or PBS, were cut with a blade in a petri dish then fixed with 1% formalin (50 mg/ml) rocking for 12 min at RT. Glycine (0.125 M) was added to stop the reaction, and the tissue was washed with PBS with serine protease inhibitor phenylmethylsulfonyl fluoride (PMSF). The tissue was resuspended in cytoplasmic lysis buffer and moved in a glass tissue grinder (Kimble Chase) for homogenisation and then filtered using a 50 μm filter (CellTrics® Sysmex). The homogenate was washed and then lysed with ChIP buffer with protease and phosphatase inhibitors. Chromatin was sonicated in ChIP buffer using a Diagenode Bioruptor® Pico (Diagenode) 30 s on/30 s off for 15 cycles. Fragment size was checked by gel electrophoresis.

For immunoprecipitation, 60 μg of chromatin were used per reaction. Sepharose protein A alone or mixed 4:1 with protein G (GE Healthcare) were used to preclear the chromatin before incubation with primary antibodies for 4 h at 4 °C. Primary antibodies used 5 μg rabbit anti-pSTAT3 (Cell Signalling #9145); 8 μg mouse anti-RUNX1 (Santa Cruz Biotechnology, Inc. #sc-365644); rabbit anti-IgG (Abcam #ab37415) and mouse anti-IgG (Cell Signalling #5415S). 20 μl of Sepharose protein A (for pSTAT3) or A/G 4:1 (for

RUNX1) were added to each sample and incubated overnight at 4 °C. Samples were collected by centrifugation and washed with low-salt wash buffer (150 mM NaCl, 20 mM Tris-HCl pH 8.1, 2 mM EDTA, 0.1% SDS, 1% Triton X-100), high-salt wash buffer (500 mM NaCl, 20 mM Tris-HCl pH 8.1, 2 mM EDTA, 0.1% SDS, 1% Triton X-100), LiCl wash buffer (10 mM Tris-HCl pH 8.1, 1 mM EDTA, 0.25 M LiCl, 1% NP-40, 1% sodium deoxycholate) and twice with TE wash buffer (10 mM Tris-HCl pH 8.1, 1 mM EDTA). Cross-links were reversed incubating with Chelex®100 resin beads (Bio-Rad #142-1253) at 99 °C for 15 min on a shaking block, and then the samples were diluted 1:1 with MQ water.

IHC

Kidneys fixed in formalin and embedded in paraffin were cut at 4 μm thickness. Sections were stained with the primary antibodies used for ChIP: rabbit anti-pSTAT3 (1:75; Cell Signalling #9145) and mouse anti-RUNX1 (1:250; Santa Cruz Biotechnology, Inc. #sc-365644). Anti-rabbit or anti-mouse Envision HRP (Dako) was used as the secondary antibody.

Renal tissue from ADPKD patients at end-stage renal failure was fixed in formalin as previously described [15]. Control tissues were obtained from donor kidneys non-suitable for transplant. All human tissue samples were collected following procedures approved by the LUMC medical ethical committee (institutional review board).

Results

Transcription factors in the PKD signature

Using a meta-analysis approach of published PKD expression profiles and in-house generated RNA-sequencing data on our *Pkd1* mutant mouse model (iKsp*Pkd1*^{del}), we recently identified 1515 genes that are commonly dysregulated across several PKD disease models, hereafter referred to as the PKD Signature [5].

We used MSigDB to identify the TFs that are part of the PKD Signature (Fig 1a). Out of the 1515 genes of the PKD Signature, we identified 92 TFs that were differentially expressed and could be involved in cyst formation and PKD development. Among the 92 TFs identified, 32 were also implicated in tissue injury/repair mechanisms based on our previously defined injury repair profile (Supplementary Table 1) [5]. Several of the herein identified TFs, such as STAT3 and MYC, are known players in ADPKD progression [17, 18]. Nevertheless, many others have never been described in ADPKD before.

Furthermore, we predicted TFs that are relevant to PKD based on the enrichment of their targets in the PKD Signature. Using the ChEA 2016 database of TF targets, we identified TFs with more experimentally verified targets (ChIP-chip or ChIP-Seq) overlapping with the PKD Signature than would be expected by chance (Fig. 1a). The TFs E2F7, TRIM28, TP63 (two different experiments in different cell lines), EGR1 and STAT3 were most significant in this analysis (Supplementary Table 2a) since targets of these TFs were mostly upregulated in PKD. Five TFs were both in the list of TFs identified based on their targets and among the 92 TFs present in the PKD Signature: EGR1, ESR1, STAT3, FOXM1 and KLF5. Thus, these TFs, as well as their identified direct targets, were dysregulated in PKD (Supplementary Table 2b). Further pathway analysis of these five TFs targets uncovered involvement in the modulation of TGF-β signalling, estrogen signalling, apoptosis, oxidative stress, interleukins signalling, adipogenesis and cellular metabolism (Supplementary Table 2c).

Validation of meta-analysis in independent samples

Our next step was to validate TFs identified in the meta-analysis in independent experimental groups of mice during PKD progression and/or the nephrotoxic injury/repair response [15]. Briefly, we induced *Pkd1* deletion in adult mice via tamoxifen administration, which leads to a slow progression of the disease. Wild-type (Wt) mice received tamoxifen as well. A week after tamoxifen administration, we injected both genotypes with 15 mg/kg of DCVC, a nephrotoxic compound or PBS as a control. At this dosage, DCVC causes a repairable renal injury that is mostly recovered 1 to 2 weeks after injection but accelerates cyst formation resulting in tubular dilations at 10 weeks and renal failure around 14 weeks of age (Supplementary Fig. 1). Mice were sacrificed at 1, 2, 5 and 10 weeks after DCVC and at kidney failure. Kidneys harvested at these time points were used to evaluate gene expression of selected TF using the Fluidigm qPCR chip (Fig. 1b). Out of the 92 TFs, 13 were selected for further analysis, based on transcript levels, altered expression in the injury/repair response and involvement in multiple molecular pathways (Supplementary Table 1). In our Fluidigm setup, we had four groups: PBS-treated Wt, DCVC-treated Wt, PBS-treated iKsp*Pkd1*^{del} and DCVC-treated iKsp*Pkd1*^{del} at five time points (1week, 2weeks, 5weeks and 10weeks after DCVC treatment and at kidney failure). Out of the 13 tested TFs, 11 were significantly different ($P < 0.05$) in PKD samples compared to Wt, while the involvement of *Irf6* and *JunB* could not be confirmed (Supplementary Table 1, Fig. 2). We also evaluated whether expression of the 13 TFs was affected by injury, by comparing DCVC versus PBS-treated animals at injury-related time points (1week, 2weeks and 5weeks after DCVC treatment). Of the 13 selected TFs, 8 were part of the previously reported injury repair profile, while 5 were not [5]. We confirmed significant injury-induced dysregulation ($P < 0.05$)

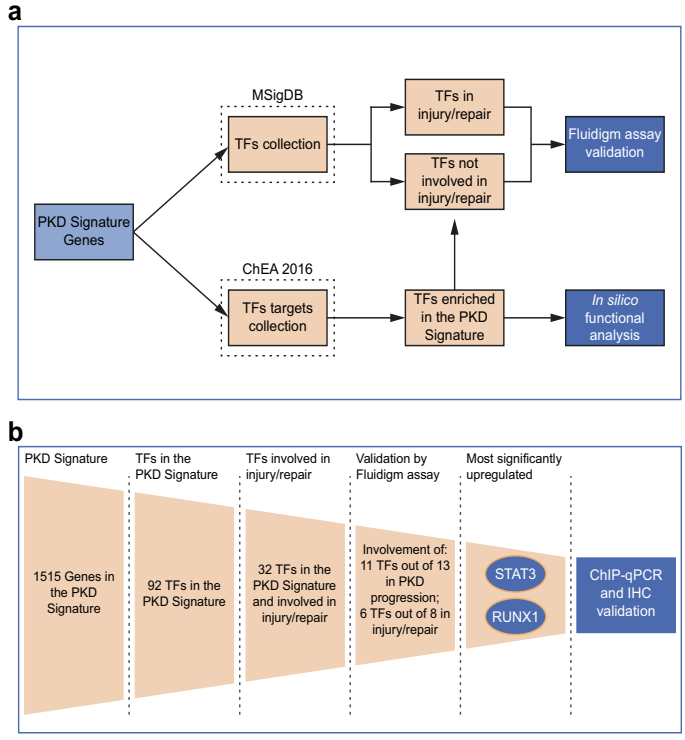

Fig. 1 Schematic representation of the workflow used for the identification and validation of TFs involved in PKD and injury/repair. **a** MSigDB was used to select the TFs in the PKD Signature. ChEA 2016 was used to select the TFs with most deregulated, experimentally verified targets in the PKD Signature (note: the ChIP-chip and ChIP-Seq experiments in ChEA 2016 were typically from cell lines not necessarily related to the kidney). The TFs identified with MSigDB in the PKD Signature were intersected with the injury signature generated in our previous work [5] to obtain TFs involved in injury/repair mechanism, and TFs involved only in PKD progression. Fluidigm assay was used to validate the expression of selected TFs identified by this analysis. The TFs identified based on their target genes using the ChEA 2016 database were intersected with the TFs identified in the PKD signature to identify the overlapping TFs. In silico pathway analysis was performed on the overlapping TFs and their target genes to identify significant pathways modulated by the TFs. **b** Schematic representation of the workflow used to identify and validate selected TFs. The two most significant TFs identified were STAT3 and RUNX1 which were further investigated in cystic kidneys using chromatin immunoprecipitation-qPCR (ChIP-qPCR) and immunohistochemistry (IHC)

of 6 out of 8 TFs predicted to be involved in the injury/repair mechanism by the meta-analysis, while we did not see any significant dysregulation of the expression of 3 out of 5 TFs that were not found in the meta-analysis (Supplementary Table 1, Fig. 2) [5]. Notably, the expression of *Runx1* and *Stat3* was most significantly affected by DCVC-induced injury and PKD progression.

Expression of two selected TFs in mouse kidneys during ADPKD progression and after injury

To further support the utility of meta-analysis approaches to new target discovery in ADPKD, we chose STAT3 and RUNX1 for additional experimental validation.

We performed immunohistochemical analysis for the active form of STAT3 (pSTAT3) and RUNX1 and studied activation and subcellular localisation. In non-injured Wt and iKsp*Pkd1*^{del} mice, pSTAT3 and RUNX1 are not detectable, except for some interstitial cells that show nuclear staining. Interestingly, after injury (at 1wk after DCVC), there was an intense nuclear expression of pSTAT3 and RUNX1 in both Wt and iKsp*Pkd1*^{del} mice (Fig. 3a and Supplementary Fig. 2a).

At 10 weeks post-DCVC, Wt mice have fully healed the renal damage and have largely pSTAT3 and RUNX1 negative kidneys, comparable to the Wt treated with PBS. Conversely, iKsp*Pkd1*^{del} mice, which already developed some mild cysts at this time point, showed expression of pSTAT3 and RUNX1 in the cyst-lining epithelial cells and some of the surrounding dilated tubules (Fig. 3b, middle panel and Supplementary Fig. 2b, middle panel). iKsp*Pkd1*^{del} mice treated with PBS, instead, have not undergone injury/repair phase nor displayed overt cyst formation at this time point and

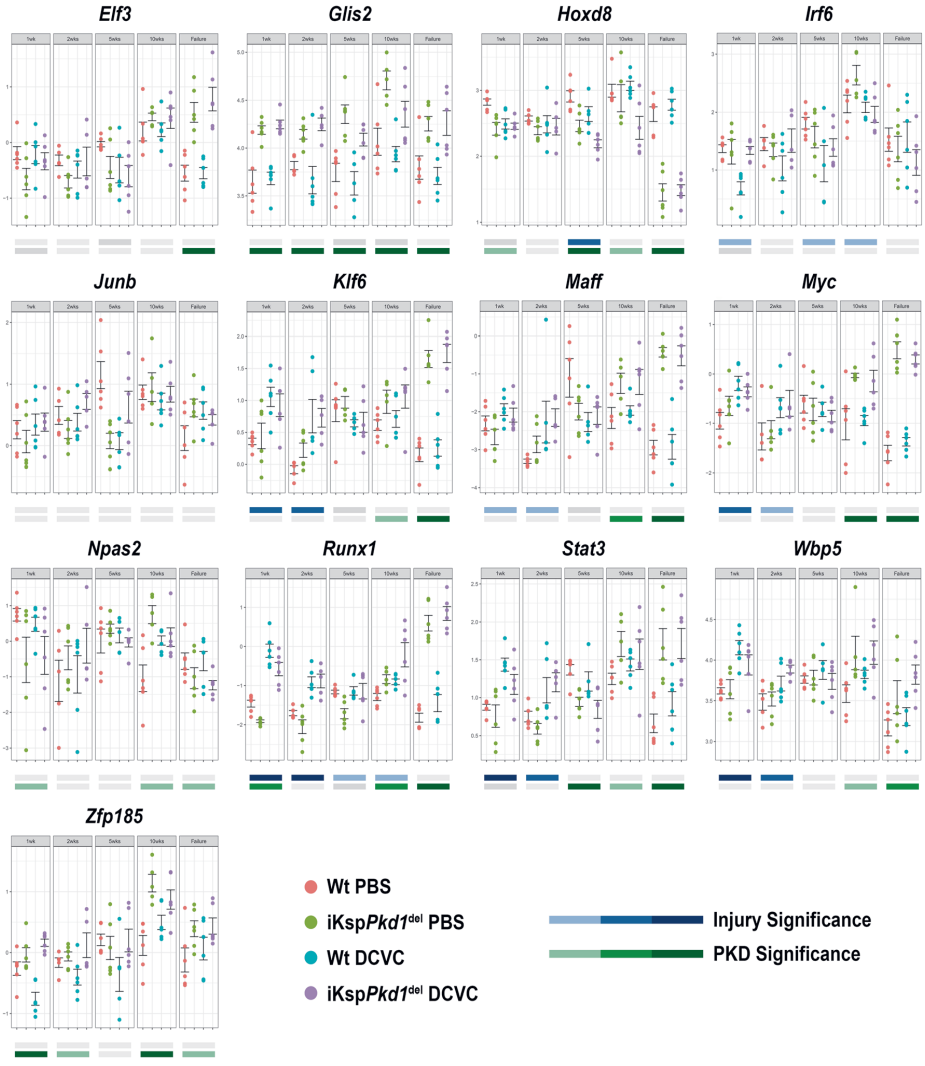

Fig. 2 Expression of selected TFs using Fluidigm assay. TFs selected from the PKD Signature for experimental validation were subjected to qRT-PCR on RNA isolated from the kidneys of iKsp*Pkd1*^{del} mice and age-matched Wt mice at 1, 2, 5 and 10 weeks after DCVC and at kidney failure. On the Y-axis, normalized Ct values (cycle threshold values) are plotted for each gene separately across the five measurement time points for four types of samples: Wt mice treated with saline (Wt PBS, salmon), iKsp*Pkd1*^{del} mice treated with saline (iKsp*Pkd1*^{del} PBS, light green), Wt mice treated with DCVC (Wt DCVC, light blue) and iKsp*Pkd1*^{del} mice treated with DCVC (iKsp*Pkd1*^{del} DCVC, light purple). The analysis was based on comparing treatment (DCVC vs PBS) and genotype (iKsp*Pkd1*^{del} vs Wt) using a two-way ANOVA test. The resulting *P* values are shown with colour codes: darkest colour shade, *P* value < 0.0005; medium colour shade, *P* value < 0.005 and low colour shade at *P* value < 0.05. *P* value ≥ 0.05 were not considered significant (grey bars). Each dot is a mouse and whiskers reflect the mean ± SD. Expression of *Glis2* and *Stat3* in Wt PBS, iKsp*Pkd1*^{del} PBS and Wt DCVC have been published in Malas et al.(2017) [5].

showed almost no expression of pSTAT3 and RUNX1, as expected.

At kidney failure, iKsp*Pkd1*^{del} mice present severe renal degeneration and cyst formation. At this time point, the expression of pSTAT3 and RUNX1 is markedly increased (Fig. 3b, right panel and Supplementary Fig. 2b, right panel). Interestingly, not only epithelial cells but also infiltrating cells stained positive for these TFs, suggesting that pSTAT3 and RUNX1 might be important in the regulation of signalling pathways in other cell types in addition to tubular epithelial cells (Fig. 3b, arrowheads).

In summary, we confirmed that pSTAT3 and RUNX1 protein expression were increased in the nuclei of tubular epithelial cells after injury and during PKD progression.

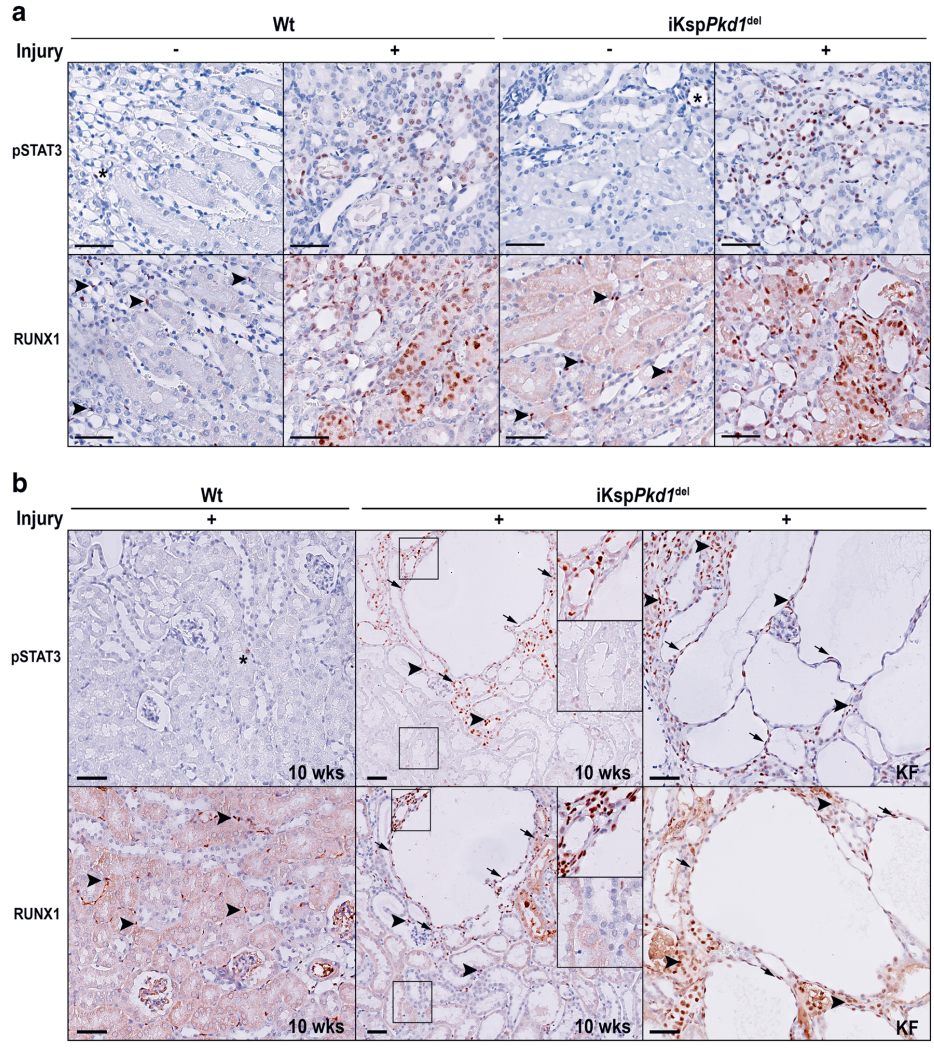

Fig. 3 Expression of pSTAT3 and RUNX1 in Wt and iKsp*Pkd1*^{del} mice after injury and during cyst progression. **a** Representative immunohistochemistry of Wt and iKsp*Pkd1*^{del} kidneys at 1 week after DCVC (+ injury) or PBS (– injury). Mice without injury showed only sporadic expression of pSTAT3 in the nuclei of tubular epithelial cells (asterisks); after injury, the expression was markedly increased both in Wt mice and in iKsp*Pkd1*^{del} mice. RUNX1 expression in non-injured kidney was present only in some interstitial cells (arrowheads); after injury, RUNX1 was visible in the nuclei of the epithelial cells. **b** Representative immunohistochemistry of Wt and iKsp*Pkd1*^{del} kidneys at 10 weeks after DCVC ("10weeks"; left and middle panel) showed expression of pSTAT3 and RUNX1 in nuclei in cyst-lining epithelia, in the epithelial cells of surrounding dilated tubules (arrows) and in infiltrating cells (arrowheads) only in cystic tissue. Expression of pSTAT3 and RUNX1 was even more increased at kidney failure ("KF"; right panel) when the kidneys are severely cystic. Scale bars 50 μm

Stat3 and Runx1 target genes were dysregulated during ADPKD progression and after injury

Although we demonstrated that pSTAT3 and RUNX1 expression were increased during ADPKD progression and after injury, both at gene and protein level, we do not know if this would translate into differences in their activity as transcriptional regulators. Thus, we quantified the expression of their target genes during PKD progression and injury/repair. To find TFs' target genes, we used the publicly available Cistrome database. For both TFs, we identified ChIP-Seq experiments and searched for peaks (targets) identified in at least two ChIP-Seq experiments. Peaks were prioritised based on (1) the number of studies they were found in, (2) their intensity levels (> 10) and (3) whether they mapped to target genes within 10 kb distance. For both TFs, the top putative target genes were crossed with the PKD Signature genes to identify targets that show differential expression in PKD. Only target

genes that were also present in the PKD Signature were selected for further analysis (Fig. 4a).

The final targets we selected are *Scp2*, *Kif22*, *Stat3* (autoregulation) and *Socs3* for STAT3 and *Runx1* (autoregulation), *Tnfrsf12a* and *Bcl3* as targets for RUNX1. We checked the expression of these targets after injury and during PKD progression in iKsp*Pkd1*^{del} and Wt mice. We found that, in iKsp*Pkd1*^{del} mice, all targets were significantly upregulated except for *Scp2*, which was downregulated, suggesting an inhibitory effect of STAT3 on *Scp2* transcription (Fig 2b, *Stat3* and *Runx1;* Fig. 4b, *Scp2*, *Kif22*, *Socs3, Tnfrsf12a* and *Bcl3*).

These data indicate that not only the level of expression of the selected TFs is dysregulated during injury/repair and PKD progression but likely also their activity, as denoted by the dysregulated expression of their target genes.

Stat3 and Runx1 ChIP-qPCR in murine renal epithelial cells

To confirm that STAT3 and RUNX1 are directly regulating the expression of the indicated target genes in the renal epithelium, we performed chromatin immunoprecipitation (ChIP) analysis followed by quantitative PCR (ChIP-qPCR). We first

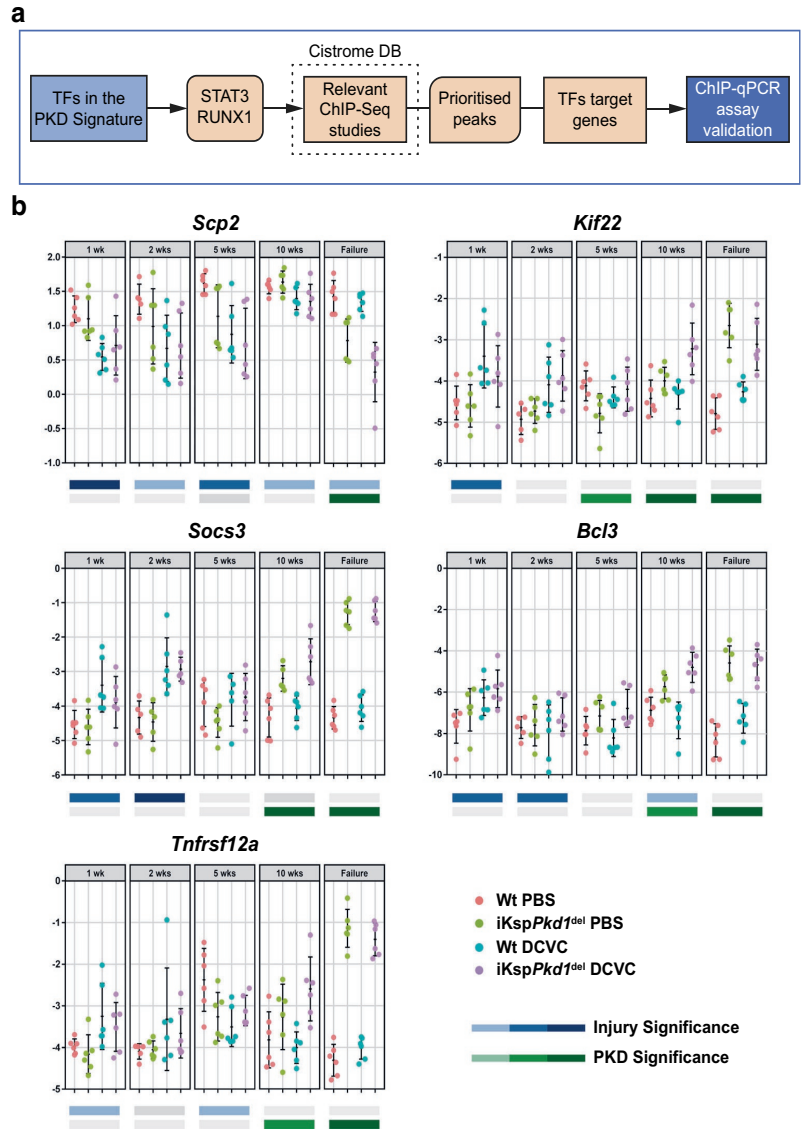

Fig. 4 Identification of STAT3 and RUNX1 target genes. **a** STAT3 and RUNX1 emerged as two leading candidates for wet-lab validation. Using Cistrome database, we identified ChIP-peaks that were used in the wet-lab validation process and led to the identification of confirmed STAT3 and RUNX1 targets. **b** Expression of STAT3 and RUNX1 targets during PKD progression. Total RNA was isolated from kidneys of Wt and iKsp*Pkd1*^{del} mice treated with PBS or DCVC at 1, 2, 5 and 10 weeks and at kidney failure. Expression of selected STAT3 (*Scp2*, *Kif22* and *Socs3*) and RUNX1 (*Bcl3*, *Tnfrsf12a*) targets was evaluated using a SYBR Green-based qPCR. On the Y-axis, normalised Ct values (cycle threshold values) are plotted. Data were analysed using a two-way ANOVA test based on comparing treatment (DCVC vs PBS) and genotype (iKsp*Pkd1*^{del} vs Wt). *P* values are reported and classified into high significance (darkest colour shade) at *P* value < 0.0005, moderate significance (medium colour shade) at *P* value < 0.005 and acceptable significance at (low colour shade) at *P* value < 0.05. *P* value ≥ 0.05 was not considered significant (grey bars). Each dot is a mouse and whiskers represent mean ± SD

confirmed that STAT3 and RUNX1 were expressed in mIMCD3 cells (Supplementary Fig 3). We then isolated chromatin and performed ChIP-qPCR. STAT3 enrichment at the promoter region of the *Scp2*, *Kif22*, *Stat3* and *Socs3* genes was significantly higher than at nonbinding regions (Fig. 5a). Also, RUNX1 showed significant enrichment at the promoter regions of its targets *Runx1*, *Tnfrsf12a* and *Bcl3* (Fig. 5b) compared to nonbinding regions.

Thus, we can conclude that STAT3 and RUNX1 are actively binding the selected target genes in renal epithelial cells.

Stat3 and Runx1 ChIP-qPCR in murine kidney tissue

We then investigated whether binding of STAT3 and RUNX1 at the promoter region of their target genes is increased in cystic kidneys compared to non-cystic kidneys.

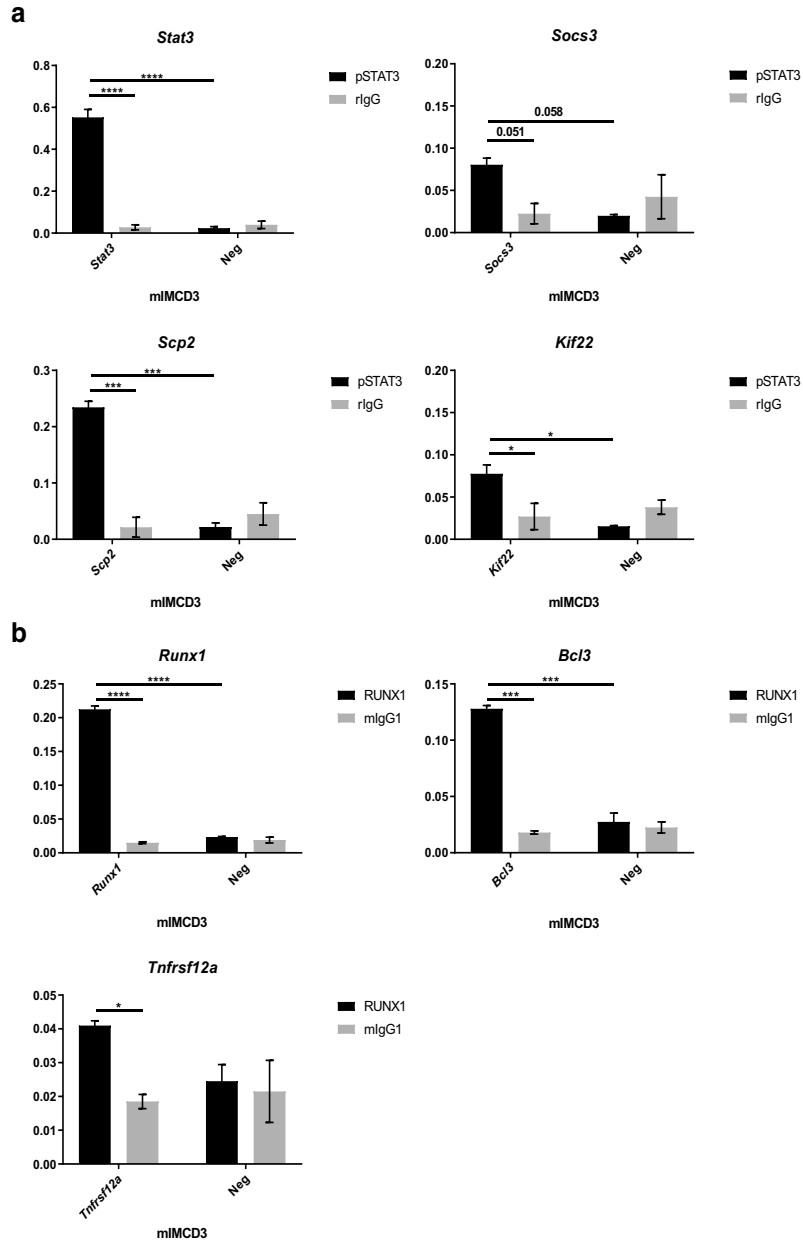

Fig. 5 ChIP validation of pSTAT3 and RUNX1 targets in mIMCD3 cells. **a** ChIP with anti-pSTAT3 antibody showed significant enrichment at the promoter region of *Scp2*, *Kif22*, *Stat3* and *Socs3* compared to a negative control antibody (rIgG) and a nonbinding region (Neg). **b** ChIP with anti-RUNX1 antibody showed a significant enrichment at the promoter region of *Runx1*, *Tnfrsf12a* and *Bcl3* compared to a negative control antibody (mIgG) and a nonbinding region (Neg). The Y-axis shows the input-normalised binding-enrichment of the TFs to the indicated genomic region. Data represent the mean of two independent ChIPs ± SD; Two-way ANOVA with Tukey's multiple comparisons test. **P* value < 0.05; ***P* value < 0.01; ****P* value < 0.001

To do so, we performed ChIP-qPCR using kidneys from iKsp*Pkd1*^{del} mice, harvested at kidney failure, as well as age- and treatment-matched Wt kidneys.

As expected, we observed a significantly increased abundance of STAT3 at *Stat3*, *Socs3*, *Scp2* and *Kif22* promoter regions in iKsp*Pkd1*^{del} mice compared to Wt (Fig. 6a, more severe iKsp*Pkd1*^{del} + DCVC and Supplementary Fig. 4a, milder iKsp*Pkd1*^{del} + PBS).

RUNX1 enrichment in iKsp*Pkd1*^{del} mice was not significantly higher than in Wt mice. However, RUNX1 enrichment was significantly higher compared to IgG at the promoter region of *Runx1* and *Bcl3* in iKsp*Pkd1*^{del} mice but not in Wt. A similar trend is observed for *Tnfrsf12a*. This means that in iKsp*Pkd1*^{del} mice, RUNX1 binding is specific, while in Wt, it is not different from the background signal. Thus, RUNX1 is actively binding its targets in cystic kidneys only (Fig. 6b,

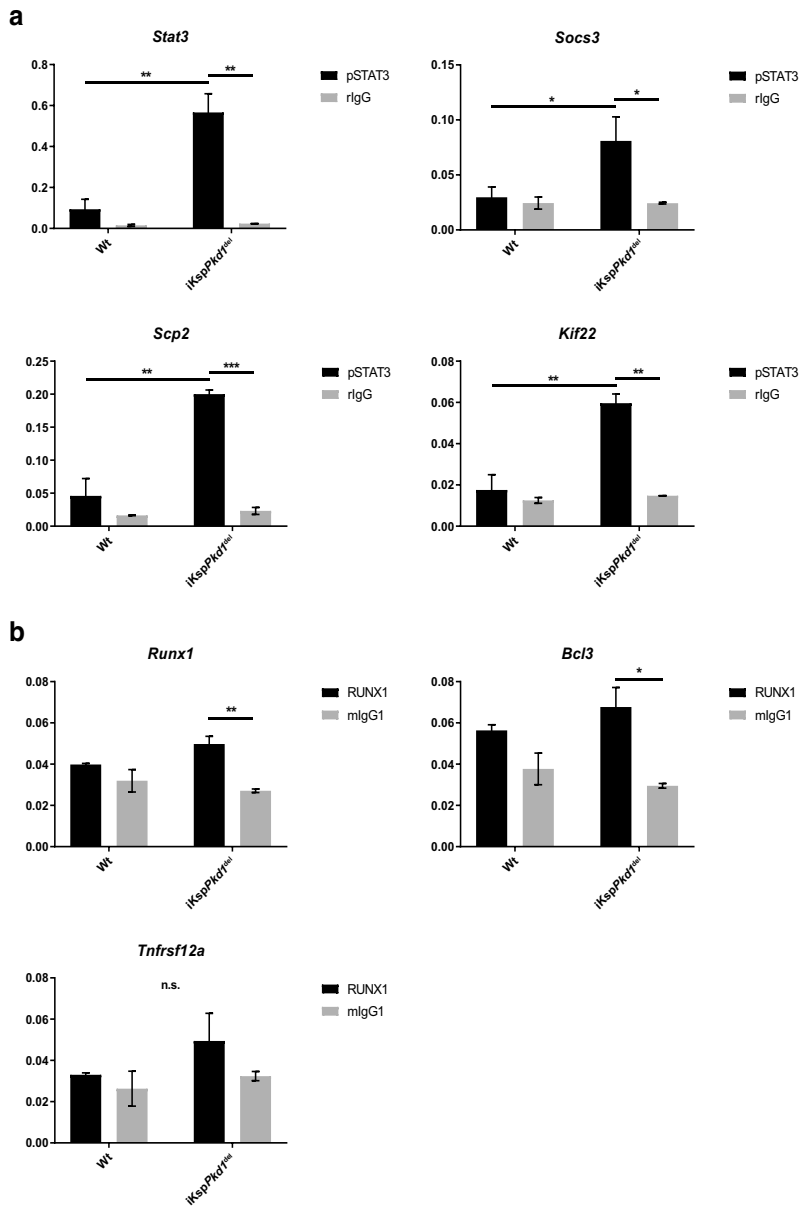

Fig. 6 Increased binding of STAT3 and RUNX1 to the promoter of target genes in cystic kidneys, shown by ChIP-qPCR. ChIP-qPCR analysis of end-stage renal disease iKsp*Pkd1*^{del} kidneys or Wt kidneys at 24 weeks after DCVC. **a** We confirmed an increased enrichment for STAT3 binding at target genes in iKsp*Pkd1*^{del} kidneys compared to Wt kidneys. **b** RUNX1 enrichment at its targets is not detected in Wt samples (no difference between RUNX1 ChIP and IgG ChIP) but detected in iKsp*Pkd1*^{del} samples. Black bars pSTAT3 or RUNX1 antibody, grey bars isotype IgG control (rIgG, rabbit IgG; mIgG, mouse IgG). The Y-axis shows the input-normalised binding-enrichment of the TFs to the indicated genomic region. Data represent the mean of two independent ChIPs ± SD; Two-way ANOVA with Tukey's multiple comparisons test. **P* value < 0.05; ***P* value < 0.01; ****P* value < 0.001

more severe iKsp*Pkd1*^{del} + DCVC and Supplementary Fig. 4b, milder iKsp*Pkd1*^{del} + PBS).

Overall, these data, in addition to the altered expression levels, show that the activity of STAT3 and RUNX1 is increased in advanced stages of PKD in mice.

Expression of TFs in kidneys of ADPKD patients

Lastly, we checked the expression of STAT3 and RUNX1 in human kidney sections obtained from ADPKD patients and healthy controls. Comparably with what was observed in mice, in healthy controls, we found only sporadic expression of pSTAT3 in the nuclei of tubular epithelial cells (Fig. 7, asterisks) and expression of RUNX1 in some infiltrating cells (Fig. 7, arrowheads). Conversely, in ADPKD patients' renal tissue, the expression of pSTAT3 and RUNX1 was increased in the nuclei of the epithelial cells and infiltrating cells (Fig. 7, right panel and Supplementary Fig. 5, right panel).

These data suggest that the TFs identified by our meta-analysis using rodent models are relevant for human ADPKD.

Discussion

Previously, we identified a list of 1515 genes dysregulated during PKD progression, which we defined as PKD Signature. We also showed a consistent overlap (about 35%) of the PKD Signature with genes normally involved in injury/repair mechanisms [5]. Now, we have put this analysis a step further by identifying and characterising TFs involved in ADPKD progression.

Using MSigDB, we identified 92 TFs in the PKD Signature and again showed that about 35% of these genes (32 out of 92) have a strong injury-related component. This is in line with a substantial body of literature indicating that injury is a significant modifier in PKD and a potential trigger of cyst formation. Indeed, renal injury causes faster cystic disease progression suggesting that events activated during the injury/repair phase are also crucial for cyst initiation and expansion [3, 4]. Moreover, cyst formation per se is a source of injury for the surrounding tissue making the two pathological processes challenging to dissect [19].

Among these 92 identified TFs, we observed known players in PKD, such as STAT3 [17, 20], c-MYC [18], SMAD2 [21], GLIS2 [22], c-JUN [23] and E2F1 [24], confirming our approach. On the other hand, we did not find TFs such as PPARα, which has been described to play a role in PKD [25]. This is likely due to the high stringency used for the definition of the PKD Signature, which allows us to get specific targets while possibly losing others [5].

Interestingly, we also identified many other TFs, never described before in PKD. Some of these TFs, such as EGR1, KLF5 and FOXM1, have been reported in literature for their involvement in injury/repair mechanisms or pathways dysregulated during PKD progression and might be interesting candidates for future studies. Indeed, *Egr1* is an early growth response gene and is downstream of the mitogen-activated protein kinase (MAPK) pathway, a pathway dysregulated in PKD [23]. EGR1 is a key regulator of proliferation, apoptosis and inflammation and was shown to be involved in renal injury and fibrosis. *Egr1* disruption protected mice from renal failure in a model of tubulointerstitial nephritis and resulted in lower activation of the TGF-β pathway [26]. Moreover, *Egr1* can be downregulated by curcumin, a compound able to reduce cyst formation in vivo [17]. Also, KLF5 was shown to play a role in renal inflammation and fibrosis since unilateral

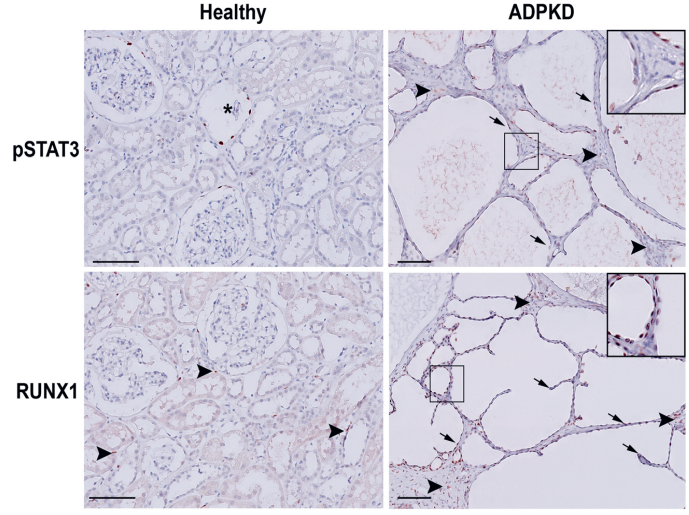

Fig. 7 pSTAT3 and RUNX1 expression in human kidneys with ADPKD. Representative immunohistochemistry of human kidneys. In healthy patients, the expression of pSTAT3 and RUNX1 was rarely detected (asterisks). In end-stage cystic kidneys from ADPKD patients, pSTAT3 and RUNX1 localised in the nuclei of the tubular epithelial cells (arrows) and infiltrating cells (arrowheads). Scale bars 100 μm

ureteral obstruction in mice haploinsufficient for *Klf5* resulted in reduced renal injury, fibrosis and infiltrating cells [27]. Thus, modulation of KLF5 activity might improve the pro-fibrotic and pro-inflammatory phenotype observed especially during the more advanced phases of PKD progression. *Foxm1* is expressed during cell proliferation and is critical for cell cycle progression. In adult tissues *Foxm1* expression is low, but after injury, its levels are dramatically increased. In particular, FOXM1 can control the expression of genes involved in the G2/M transition phase. Cell-cycle arrest in G2/M phase is associated with pro-fibrotic cytokines production by proximal tubular cells [28]. Not surprisingly, these three TFs are involved in PKD since aberrant extracellular matrix (ECM) deposition is commonly found in PKD patients and animal models of PKD, not only in ESRD but also in early stage [29]. This suggests that increased ECM deposition may be contributing to cyst formation and not barely be a consequence of it, as shown for laminin-alpha5 [30] and integrins-beta1 [31], which mutation could affect the cystic phenotype. Thus, modulation of pro-fibrotic processes could be a valuable strategy to modulate PKD progression.

EGR1, KLF5 and FOXM1, together with ESR1 and STAT3, were also among the significantly enriched PKD Signature TFs identified based on their target genes annotated in the ChEA 2016 database. Pathway analysis of the targets of these TFs, using Genetrail2 and Wikipathways, revealed enrichment for pathways known to play a role in PKD progression, such as the TGF-β pathway, oxidative stress, cellular metabolism, interleukins signalling, adipogenesis, estrogen signalling and apoptosis [21, 32–35]. Using this approach, we also identified TFs not directly present in the PKD Signature. Interestingly, the top five TFs identified based on their targets were all described in literature to be involved in the progression of PKD (STAT3)[17, 20, 36] or in processes relevant for PKD like angiogenesis (E2F7)[37], DNA damage response (E2F7, TRIM28)[38, 39], renal injury and fibrosis (EGR1)[26], epithelial cell proliferation, apoptosis and adhesion (TP63)[40]. Nevertheless, apart from STAT3, the TFs themselves had never been associated with PKD before and therefore could be interesting subjects for future studies. Surprisingly, we did not find back RUNX1 in this list as the level of enrichment was just below the significance threshold (data not shown). Nevertheless, we confirmed increased expression and activity of RUNX1 during PKD progression in mice and human ADPKD kidneys. Thus, we speculate that the absence of RUNX1, as well as other TFs potentially involved in PKD, is due to limitations related with the ChEA database, such as the source of ChIP-data, the way the different studies have been analysed and the actual TFs included in the database.

To further test and validate our approach, we selected for additional wet-lab validation STAT3 and RUNX1 as they showed the most significant change in expression both in PKD progression and injury. By performing ChIP-qPCR for STAT3 and RUNX1 in ADPKD-affected kidneys, we confirmed increased transcriptional activity in cystic kidneys for these TFs. Persistent activation of STAT3 has been described in several mouse models for ADPKD as well as in human cystic tissues [17, 20, 36]. STAT3 usually is not active in adult kidneys but is abundantly present, suggesting that it can be readily activated at needs, such as after injury [36]. Indeed, STAT3 activation has been shown in several different mouse models with renal injury [41, 42]. Thus, the fact that we found back STAT3 and several of its putative targets in our signature proved the reliability of our meta-analysis.

RUNX1 involvement in ADPKD has never been described before. RUNX1 is one of the Runt domain TFs, together with RUNX2 and RUNX3. RUNX2 expression has been shown to be regulated by PC1 in osteoblasts, proving the existence of an interaction between the two proteins [43]. Nevertheless, expression of RUNX2 or RUNX3 is not increased after injury nor during disease progression in murine (cystic) kidneys (RNA-Seq data identifier E-MTAB-5319 published in Malas et al., 2017 [5]). RUNX1 is expressed in the epithelium of several organs during development, among which the kidneys [44]. It participates in the regulation of cell cycle, cell proliferation and apoptosis [45] and has been described in several models for lung, muscle and brain injury [46–48]. Recently, a study was published suggesting that RUNX1 is an important regulator of TGF-β-induced renal tubular epithelial-to-mesenchymal transition (EMT) and fibrosis [49]. As mentioned above, TGF-β signalling is involved in ECM deposition and cyst progression and is partly responsible for the EMT observed in cystic kidneys. Modulation of TGF-β-related signalling is associated with amelioration of the cystic phenotype [21]. Thus, it is plausible that RUNX1 might play a role in ADPKD progression. In fact, inhibition of STAT3 signalling with more or less specific inhibitors, such as curcumin, pyrimethamine and S3I-201, has been proven to improve the cystic phenotype in different mouse models [17, 20, 36]. Similarly, we propose that targeting RUNX1, for example, using microRNAs as described for prostate cancer [50], or other molecular or pharmacological approaches, might also result in amelioration of the cystic phenotype.

We observed increased expression of STAT3 and RUNX1 also after injury in Wt mice, suggesting that these TFs orchestrate injury/repair mechanisms and that increased expression is not necessarily related to *Pkd1* deletion. Notably, dissecting PKD progression and injury is not easy, since injury can speed up cyst initiation/growth, which in turn causes injury to the surrounding tissue. Therefore, it is plausible that both STAT3 and RUNX1 are facilitating PKD progression by activating injury/repair pathways normally inactive in fully developed and healthy kidneys.

To conclude, our comprehensive analyses identified a signature of TFs differentially expressed in PKD and to a certain

extent also in injury/repair. Several of these TFs are involved in processes able to support cyst formation and progression, nevertheless were never described before in PKD, suggesting that they might be interesting targets for therapy. However, further analyses are needed to identify the molecular pathways that these TFs modulate to contribute to PKD progression and cyst formation. Additionally, the TFs we identified are a subset of the TFs involved in PKD and not a comprehensive list. This is due to limitations in the annotation databases we used and RNA-Seq technology. To establish a comprehensive list of TFs involved in PKD and/or injury, further studies must be conducted on protein levels and protein phosphorylation status. That said, our approach was capable of robustly identifying 92 TFs, and additional wet-lab validations confirmed the involvement of RUNX1 and STAT3 making this paper a starting point to understand the role of TFs in PKD progression.

Author contribution C.F., T.B.M., P.A.t.H. and D.J.M.P. conceived and designed research; C.F. performed experiments; L.V. performed immunohistochemistry; C.F., T.B.M. and J.B. analysed data; C.F., T.B.M., P.A.t.H. and D.J.M.P. interpreted results of experiments; C.F. and T.B.M. prepared figures; C.F. drafted manuscript; C.F., T.B.M., J.B., P.A.t.H. and D.J.M.P. edited and revised manuscript.

Funding information This work was supported by grants from the People Program (Marie Curie Actions) of the European Union's Seventh Framework Program FP7/2077-2013 under Research Executive Agency Grant Agreement 317246.

Compliance with ethical standards

Conflict of interest The authors declare that they have no competing interests.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

1. Harris PC, Torres VE (2014) Genetic mechanisms and signaling pathways in autosomal dominant polycystic kidney disease. J Clin Invest 124(6):2315–2324
2. Leonhard WN, Happe H, Peters DJ (2016) Variable cyst development in autosomal dominant polycystic kidney disease: the biologic context. J Am Soc Nephrol 27(12):3530–3538
3. Patel V, Li L, Cobo-Stark P, Shao X, Somlo S, Lin F, Igarashi P (2008) Acute kidney injury and aberrant planar cell polarity induce cyst formation in mice lacking renal cilia. Hum Mol Genet 17(11): 1578–1590
4. Happe H, Leonhard WN, van der Wal A, van de Water B, Lantinga-van Leeuwen IS, Breuning MH, de Heer E, Peters DJ (2009) Toxic tubular injury in kidneys from Pkd1-deletion mice accelerates cystogenesis accompanied by dysregulated planar cell polarity and canonical Wnt signaling pathways. Hum Mol Genet 18(14): 2532–2542
5. Malas TB, Formica C, Leonhard WN, Rao P, Granchi Z, Roos M, Peters DJ, t Hoen PA (2017) Meta-analysis of polycystic kidney disease expression profiles defines strong involvement of injury repair processes. Am J Physiol Ren Physiol 312(4):F806–F817
6. Messina DN, Glasscock J, Gish W, Lovett M (2004) An ORFeome-based analysis of human transcription factor genes and the construction of a microarray to interrogate their expression. Genome Res 14(10b):2041–2047
7. Moreland RT, Ryan JF, Pan C, Baxevanis AD (2009) The homeodomain resource: a comprehensive collection of sequence, structure, interaction, genomic and functional information on the homeodomain protein family. Database (Oxford) 2009:bap004
8. Lachmann A, Xu H, Krishnan J, Berger SI, Mazloom AR, Ma'ayan A (2010) ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. Bioinformatics 26(19): 2438–2444
9. Stockel D, Kehl T, Trampert P, Schneider L, Backes C, Ludwig N, Gerasch A, Kaufmann M, Gessler M, Graf N, Meese E, Keller A, Lenhof HP (2016) Multi-omics enrichment analysis using the GeneTrail2 web service. Bioinformatics 32(10):1502–1508
10. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 43(7):e47
11. Mei S, Qin Q, Wu Q, Sun H, Zheng R, Zang C, Zhu M, Wu J, Shi X, Taing L, Liu T, Brown M, Meyer CA, Liu XS (2017) Cistrome data browser: a data portal for ChIP-Seq and chromatin accessibility data in human and mouse. Nucleic Acids Res 45(D1):D658–D662
12. Karolchik D, Hinrichs AS, Kent WJ (2009) The UCSC genome browser. Curr Protoc Bioinformatics Chapter 1:Unit1 4. https://doi.org/10.1002/0471250953.bi0104s28
13. Lantinga-van Leeuwen IS, Dauwerse JG, Baelde HJ, Leonhard WN, van de Wal A, Ward CJ, Verbeek S, DeRuiter MC, Breuning MH, de Heer E, Peters DJM (2004) Lowering of Pkd1 expression is sufficient to cause polycystic kidney disease. Hum Mol Genet 13(24):3069–3077
14. Lantinga-van Leeuwen IS, Leonhard WN, van der Wal A, Breuning MH, de Heer E, Peters DJ (2007) Kidney-specific inactivation of the Pkd1 gene induces rapid cyst formation in developing kidneys and a slow onset of disease in adult mice. Hum Mol Genet 16(24): 3188–3196
15. Formica C, Happe H, Veraar KAM, Vortkamp A, Scharpfenecker M, McNeill H, Peters DJM (2019) Four-jointed knock-out delays renal failure in an ADPKD model with kidney injury. J Pathol. https://doi.org/10.1002/path.5286
16. Nelson JD, Denisenko O, Bomsztyk K (2006) Protocol for the fast chromatin immunoprecipitation (ChIP) method. Nat Protoc 1(1): 179–185
17. Leonhard WN, van der Wal A, Novalic Z, Kunnen SJ, Gansevoort RT, Breuning MH, de Heer E, Peters DJ (2011) Curcumin inhibits cystogenesis by simultaneous interference of multiple signaling pathways: in vivo evidence from a Pkd1-deletion model. Am J Physiol Ren Physiol 300(5):F1193–F1202
18. Trudel M, Dagati V, Costantini F (1991) C-Myc as an inducer of polycystic kidney-disease in transgenic mice. Kidney Int 39(4): 665–671
19. Leonhard WN, Zandbergen M, Veraar K, van den Berg S, van der Weerd L, Breuning M, de Heer E, Peters DJM (2015) Scattered deletion of PKD1 in kidneys causes a cystic snowball effect and recapitulates polycystic kidney disease. J Am Soc Nephrol 26(6): 1322–1333
20. Takakura A, Nelson EA, Haque N, Humphreys BD, Zandi-Nejad K, Frank DA, Zhou J (2011) Pyrimethamine inhibits adult

polycystic kidney disease by modulating STAT signaling pathways. Hum Mol Genet 20(21):4143–4154
21. Hassane S, Leonhard WN, van der Wal A, Hawinkels LJ, Lantinga-van Leeuwen IS, ten Dijke P, Breuning MH, de Heer E, Peters DJ (2010) Elevated TGFbeta-Smad signalling in experimental Pkd1 models and human patients with polycystic kidney disease. J Pathol 222(1):21–31
22. Lu D, Rauhauser A, Li B, Ren C, McEnery K, Zhu J, Chaki M, Vadnagara K, Elhadi S, Jetten AM, Igarashi P, Attanasio M (2016) Loss of Glis2/NPHP7 causes kidney epithelial cell senescence and suppresses cyst growth in the Kif3a mouse model of cystic kidney disease. Kidney Int 89(6):1307–1323
23. Le NH, van der Wal A, van der Bent P, Lantinga-van Leeuwen IS, Breuning MH, van Dam H, de Heer E, Peters DJ (2005) Increased activity of activator protein-1 transcription factor components ATF2, c-Jun, and c-Fos in human and mouse autosomal dominant polycystic kidney disease. J Am Soc Nephrol 16(9):2724–2731
24. Zhou X, Fan LX, Sweeney WE Jr, Denu JM, Avner ED, Li X (2013) Sirtuin 1 inhibition delays cyst formation in autosomal-dominant polycystic kidney disease. J Clin Invest 123(7):3084–3098
25. Lakhia R, Yheskel M, Flaten A, Quittner-Strom EB, Holland WL, Patel V (2018) PPAR alpha agonist fenofibrate enhances fatty acid beta-oxidation and attenuates polycystic kidney and liver disease in mice. Am J Physiol Ren Physiol 314(1):F122–F131
26. Ho LC, Sung JM, Shen YT, Jheng HF, Chen SH, Tsai PJ, Tsai YS (2016) Egr-1 deficiency protects from renal inflammation and fibrosis. J Mol Med (Berl) 94(8):933–942
27. Fujiu K, Manabe I, Nagai R (2011) Renal collecting duct epithelial cells regulate inflammation in tubulointerstitial damage in mice. J Clin Invest 121(9):3425–3441
28. 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(5):535–543, 531p following 143
29. Song CJ, Zimmerman KA, Henke SJ, Yoder BK (2017) Inflammation and fibrosis in polycystic kidney disease. Results Probl Cell Differ 60:323–344
30. Shannon MB, Patton BL, Harvey SJ, Miner JH (2006) A hypomorphic mutation in the mouse laminin alpha5 gene causes polycystic kidney disease. J Am Soc Nephrol 17(7):1913–1922
31. Lee K, Boctor S, Barisoni LM, Gusella GL (2015) Inactivation of integrin-beta1 prevents the development of polycystic kidney disease after the loss of polycystin-1. J Am Soc Nephrol 26(4):888–895
32. Tao YX, Zafar I, Kim J, Schrier RW, Edelstein CL (2008) Caspase-3 gene deletion prolongs survival in polycystic kidney disease. J Am Soc Nephrol 19(4):749–755
33. Padovano V, Podrini C, Boletta A, Caplan MJ (2018) Metabolism and mitochondria in polycystic kidney disease research and therapy. Nat Rev Nephrol 14(11):678–687
34. Merta M, Tesar V, Zima T, Jirsa M, Rysava R, Zabka J (1997) Cytokine profile in autosomal dominant polycystic kidney disease. Biochem Mol Biol Int 41(3):619–624
35. Stringer KD, Komers R, Osman SA, Oyama TT, Lindsley JN, Anderson S (2005) Gender hormones and the progression of experimental polycystic kidney disease. Kidney Int 68(4):1729–1739
36. Talbot JJ, Shillingford JM, Vasanth S, Doerr N, Mukherjee S, Kinter MT, Watnick T, Weimbs T (2011) Polycystin-1 regulates STAT activity by a dual mechanism. Proc Natl Acad Sci U S A 108(19):7985–7990
37. Weijts BG, Bakker WJ, Cornelissen PW, Liang KH, Schaftenaar FH, Westendorp B, de Wolf CA, Paciejewska M, Scheele CL, Kent L, Leone G, Schulte-Merker S, de Bruin A (2012) E2F7 and E2F8 promote angiogenesis through transcriptional activation of VEGFA in cooperation with HIF1. EMBO J 31(19):3871–3884
38. Carvajal LA, Hamard P-J, Tonnessen C, Manfredi JJ (2012) E2F7, a novel target, is up-regulated by p53 and mediates DNA damage-dependent transcriptional repression. Genes Dev 26(14):1533–1545
39. Iyengar S, Farnham PJ (2011) KAP1 protein: an enigmatic master regulator of the genome. J Biol Chem 286(30):26267–26276
40. Carroll DK, Brugge JS, Attardi LD (2007) p63, cell adhesion and survival. Cell cycle (Georgetown, Tex) 6(3):255–261
41. Liu J, Zhong Y, Liu G, Zhang X, Xiao B, Huang S, Liu H, He L (2017) Role of Stat3 signaling in control of EMT of tubular epithelial cells during renal fibrosis. Cell Physiol Biochem 42(6):2552–2558
42. Nechemia-Arbely Y, Barkan D, Pizov G, Shriki A, Rose-John S, Galun E, Axelrod JH (2008) IL-6/IL-6R axis plays a critical role in acute kidney injury. J Am Soc Nephrol 19(6):1106–1115
43. Xiao Z, Zhang S, Mahlios J, Zhou G, Magenheimer BS, Guo D, Dallas SL, Maser R, Calvet JP, Bonewald L, Quarles LD (2006) Cilia-like structures and polycystin-1 in osteoblasts/osteocytes and associated abnormalities in skeletogenesis and Runx2 expression. J Biol Chem 281(41):30884–30895
44. Pozner A, Lotem J, Xiao C, Goldenberg D, Brenner O, Negreanu V, Levanon D, Groner Y (2007) Developmentally regulated promoter-switch transcriptionally controls Runx1 function during embryonic hematopoiesis. BMC Dev Biol 7:84
45. Zhang L, Fried FB, Guo H, Friedman AD (2008) Cyclin-dependent kinase phosphorylation of RUNX1/AML1 on 3 sites increases transactivation potency and stimulates cell proliferation. Blood 111(3):1193–1200
46. Tang X, Sun L, Jin X, Chen Y, Zhu H, Liang Y, Wu Q, Han X, Liang J, Liu X, Liang Z, Wang G, Luo F (2017) Runt-related transcription factor 1 regulates LPS-induced acute lung injury via NF-kappaB signaling. Am J Respir Cell Mol Biol 57(2):174–183
47. Umansky KB, Gruenbaum-Cohen Y, Tsoory M, Feldmesser E, Goldenberg D, Brenner O, Groner Y (2015) Runx1 transcription factor is required for myoblasts proliferation during muscle regeneration. PLoS Genet 11(8):e1005457
48. Logan TT, Villapol S, Symes AJ (2013) TGF-beta superfamily gene expression and induction of the Runx1 transcription factor in adult neurogenic regions after brain injury. PLoS One 8(3):e59250
49. Zhou T, Luo M, Cai W, Zhou S, Feng D, Xu C, Wang H (2018) Runt-related transcription factor 1 (RUNX1) promotes TGF-beta-induced renal tubular epithelial-to-mesenchymal transition (EMT) and renal fibrosis through the PI3K subunit p110delta. EBioMedicine 31:217–225
50. Zhang G, Han G, Zhang X, Yu Q, Li Z, Li Z, Li J (2018) Long non-coding RNA FENDRR reduces prostate cancer malignancy by competitively binding miR-18a-5p with RUNX1. Biomarkers 23(5):435–445

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

CHAPTER 4
PRIORITIZATION OF NOVEL ADPKD DRUG CANDIDATES FROM DISEASE-STAGE SPECIFIC GENE EXPRESSION PROFILES

Tareq B. Malas, Wouter N. Leonhard, Hester Bange, Zoraide Granchi, Kristina M.Hettne, Gerard J.P. Van Westen, Leo S. Price, Peter A.C.'t Hoen, Dorien J.M. Peters

EBioMedicine, Volume 51, 2020, 102585, ISSN 2352-3964, https://doi.org/10.1016/j.ebiom.2019.11.046.

December 2019

Prioritization of novel ADPKD drug candidates from disease-stage specific gene expression profiles

Tareq B. Malas[a], Wouter N. Leonhard[a], Hester Bange[b], Zoraide Granchi[c], Kristina M. Hettne[a], Gerard J.P. Van Westen[d], Leo S. Price[b], Peter A.C. 't Hoen[a,e,1], Dorien J.M. Peters[a,*,1]

[a] Department of Human Genetics, Leiden University Medical Center, Leiden, the Netherlands
[b] OcellO B.V., Leiden, the Netherlands
[c] GenomeScan B.V., Plesmanlaan 1/D, 2333 BZ Leiden, the Netherlands
[d] Drug Discovery and Safety, Leiden Academic Center for Drug Research, Einsteinweg 55, 2333 CC, Leiden, the Netherlands
[e] Center for Molecular and Biomolecular Informatics, Radboud Institute for Molecular Life Sciences, Radboud University Medical Center Nijmegen, Nijmegen, the Netherlands

ARTICLE INFO

Article History:
Received 13 September 2019
Revised 14 November 2019
Accepted 26 November 2019
Available online 24 December 2019

Keywords:
Drug repurposing
Autosomal dominant polycystic kidney disease
RNA-Sequencing
3D cyst assay
Cheminformatics

ABSTRACT

Background: Autosomal Dominant Polycystic Kidney Disease (ADPKD) is one of the most common causes of end-stage renal failure, caused by mutations in *PKD1* or *PKD2* genes. Tolvaptan, the only drug approved for ADPKD treatment, results in serious side-effects, warranting the need for novel drugs.
Methods: In this study, we applied RNA-sequencing of *Pkd1cko* mice at different disease stages, and with/without drug treatment to identify genes involved in ADPKD progression that were further used to identify novel drug candidates for ADPKD. We followed an integrative computational approach using a combination of gene expression profiling, bioinformatics and cheminformatics data.
Findings: We identified 1162 genes that had a normalized expression after treating the mice with drugs proven effective in preclinical models. Intersecting these genes with target affinity profiles for clinically-approved drugs in ChEMBL, resulted in the identification of 116 drugs targeting 29 proteins, of which several are previously linked to Polycystic Kidney Disease such as Rosiglitazone. Further testing the efficacy of six candidate drugs for inhibition of cyst swelling using a human 3D-cyst assay, revealed that three of the six had cyst-growth reducing effects with limited toxicity.
Interpretation: Our data further establishes drug repurposing as a robust drug discovery method, with three promising drug candidates identified for ADPKD treatment (Meclofenamic Acid, Gamolenic Acid and Birinapant). Our strategy that combines multiple-omics data, can be extended for ADPKD and other diseases in the future.
Funding: European Union's Seventh Framework Program, Dutch Technology Foundation Stichting Technische Wetenschappen and the Dutch Kidney Foundation.
© 2019 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license. (http://creativecommons.org/licenses/by/4.0/)

1. Introduction

Drug repurposing, defined as the application of known drugs and compounds to treat new indications, is seen as a bypass for the long and expensive process of developing new drugs. Estimates show that drug repurposing can save more than 50% of the cost and time needed to develop new drugs [1]. In the past, accidental discovery, unintended side effects or obvious follow-on indications have led to new uses of such drugs [2]. Notable examples of drug repurposing include, Minoxidil (originally tested for hypertension; now indicated for hair loss) and Viagra (originally tested for angina; now indicated for erectile dysfunction and pulmonary hypertension). Current drug repurposing efforts span the spectrum from blind screening chemical libraries against specific cell lines [3] or against cellular organisms [4], to serial testing in animal models [5], and to data-driven computational methods [6]. The latter category explores the fact that a single molecule can act on several targets, making it valuable to indications where these targets are also relevant [7]. Gene expression profiles generated with expression microarrays or RNA-sequencing, have been used for the identification of druggable targets and pathways [8–10] and are suited for the identification of drug repurposing candidates under the assumption that diseases that share aberrant molecular processes may be targeted by the same drugs. However, gene expression profiles have mainly be used in isolation and

* Corresponding author: Dorien J.M. Peters, Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands.
E-mail address: d.j.m.peters@lumc.nl (D.J.M. Peters).
[1] These authors contributed equally.

https://doi.org/10.1016/j.ebiom.2019.11.046
2352-3964/© 2019 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license. (http://creativecommons.org/licenses/by/4.0/)

Research in context

Evidence before this study

Autosomal Dominant Polycystic Kidney Disease (ADPKD) is a progressive kidney disease, with 50% of patients reaching end-stage kidney disease at the age of 55. Fluid-filled cysts that gradually replace normal kidney parenchyma, accompanied by massive fibrosis, are identified as the main cause of renal failure. Tolvaptan is currently the only approved drug for ADPKD treatment, but with serious side-effects (i.e. diuresis). Therefore, there is a need for drugs that specifically target the formation and growth of cysts, to slow down or halt disease progression.

Added value of this study

Using a novel approach that combines bio and chemo-informatics, we repurpose drugs for the treatment of ADPKD. We compared transcriptomic data of ADPKD mouse models at different disease stages, as well as before and after drug treatment, to identify genes that are involved in ADPKD progression. By screening the ChEMBL drug-protein interaction database, we prioritized a list of candidate drugs that target ADPKD progression-associated genes. Finally, we showed that three out of six selected candidate compounds exhibit cyst-growth reducing effects in vitro, without toxic effects.

Implications of all the available evidence

We have identified three novel compounds that could be further investigated and developed for the treatment of ADPKD, these are Meclofenamic Acid, Gamolenic Acid, and Birinapant. Furthermore, our approach is applicable to other diseases, provided that high quality transcriptomic/proteomics data is available for integration with large scale drug affinity and activity data.

integrative approaches where gene expression profiles are combined with other information are scarce.

Here we have undertaken a novel approach to repurpose drugs for the treatment of Autosomal Dominant Polycystic Kidney Disease (ADPKD). ADPKD is a genetic disease of the kidney, with a prevalence of 4 to 10/10,000, it is one of the most common causes of end-stage renal failure [11,12]. ADPKD is characterized by the gradual replacement of normal kidney parenchyma by fluid-filled cysts and fibrotic tissue with age, ultimately leading to end-stage renal disease in most patients. The main genes mutated in patients with ADPKD are the *PKD1* and *PKD2* genes [13]. ADPKD shows variable disease progression, with 50% of patients developing end-stage kidney disease by the age of 60. While advances have been made in slowing the progression of some other forms of chronic kidney disease, standard treatments have not reduced the need for renal replacement therapy in ADPKD [14,15]. Unfortunately, several experimental interventions have recently failed to show significant benefit in slowing the rate of functional decline [16–18], while the interventions with positive outcomes, including the approved drug Tolvaptan, reported modest effects [19,20].

The difficulty in identifying drugs for ADPKD treatment can be partially attributed to the lack of understanding of the functions of the *PKD1* and *PKD2*-gene products, and on how their inactivation leads to cyst development. Strategies are focused on therapies that can slow the rate of disease progression in PKD patients. The identification of more and better drugs would require a macro-level understanding of the key molecular pathways contributing to cyst initiation and growth in patients and animal models. Transcriptomics deep-sequencing of disease states was proven successful in identifying promising drug candidates in several examples [21,22].

By sequencing mild, moderate and advanced stages of ADPKD mouse models, we identified genes involved in ADPKD progression. To further validate these genes involvement in disease progression, we compared their expression to the expression profiles of drug-treated ADPKD mouse models and looked-for gene expression alterations that are normalized after drug treatment. These genes have been included in a drug repurposing analysis in which targets of drugs published in ChEMBL have been compared to our expression profiles. This resulted in the identification of several drugs that potentially can be repurposed for ADPKD. We validated several of these compounds in a 3D cyst culture assay and propose them as potential candidates for ADPKD treatment (Supplementary Figure 1).

2. Materials and methods

2.1. Animal models and drug treatments

2.1.1. Mice used in the ADPKD progression analysis

The inducible kidney-specific *Pkd1*-deletion mouse model (tam-KspCad-CreER^{T2};*Pkd1*$^{lox2\text{-}11;lox2\text{-}11}$, referred to as iKsp-*Pkd1*^{del}) and tamoxifen treatments have previously been described [23]. In this study mutant mice are called *Pkd1cko* mice. RNA sequencing was done on kidneys from 5 adult Wild-type (Wt) mice and 24 *iKsp-Pkd1*^{del} mice with tamoxifen-induced gene disruption at the age of 38 or approximately 90 days (Mutant). Four mice per group were sacrificed at 2wk, 3wk and 6wks after tamoxifen administration. Five mice were sacrificed at 11wk of age, 4 at 12wk of age and 3 mice at 15wk after tamoxifen administration (Supplementary Table 1, Supplementary Figure 2). In addition, a young PKD model was analyzed with tamoxifen treatment at postnatal age of 10 days, as previously described [24], and the kidneys were harvested at age of 4.7 weeks (n = 3). Blood sampling and blood urea measurements were performed using Reflotron technology (Kerkhof Medical Service) as described previously [25]. Only male mice were used.

2.1.2. Ethics statement

All the animal experiments were evaluated and approved by the local animal experimental committee of the Leiden University Medical Center (LUMC) and the Commission Biotechnology in Animals of the Dutch Ministry of Agriculture.

2.1.3. Drug treated mice

Rapamycin (Sirolimus), Curcumin and soluble activin receptor IIB Fc (sActRIIB-Fc) treated *Pkd1cko* mice and controls were previously published [23,24,26] (Supplementary Figure 2).

2.1.4. Measurement of disease progression in ADPKD model

2KW/BW was used as measurement for disease severity and strongly correlated with the cystic index (Supplementary Figure 3).

2.2. Statistical analysis

2.2.1. Processing of RNA sequencing samples

RNA sequencing was performed on the Illumina® HiSeq 2500. The Illumina® mRNA-Seq Sample Prep Kit was used to process the sample according the Illumina protocol "Preparing Samples for Sequencing of mRNA" (1,004,898 Rev. D). Briefly, mRNA was isolated from total RNA using the oligodT magnetic beads. After fragmentation of the mRNA, a cDNA synthesis was performed. This was used for ligation with the sequencing adapters and PCR amplification of the resulting product. The quality and yield after sample preparation were measured with a DNA 1000 Lab-on-a-Chip (Agilent Technologies). The size of the resulting products was consistent

with the expected size distribution (a broad peak between 300−500 bp on a DNA 1000 chip). Clustering and DNA sequencing using the Illumina cBot and HiSeq 2500 was performed according to manufacturer's protocols. A concentration of 15.0 pM of DNA was used. Detailed run information per group is provided in Supplementary Table 1. HiSeq control software HCS v2.2.38 was used. Image analysis, base calling, and quality check was performed with the Illumina data analysis pipeline RTA v1.18.64 and Bcl2fastq v1.8.4. All samples had a quality score Q30 for more than 93.6% of reads. Resulting reads were aligned to the mouse reference genome version GRCm38 using Tophat v.2.0.12 with default parameters [27]. The only exception is the use of the no-coverage-search which does not perform an initial coverage search against the genome, thus reducing substantially the computational time. After alignment, HTSeq-count (Version 0.6.1) was used to estimate gene expression by counting reads that were mapped to the reference genome GRCm38 exons of each gene using the following options: -s (stranded) = no, -a (mapping quality) = 10, -m (mode) = intersection-nonempty, -i (identification) = gene_id -t (feature to count) = exon. Gene counts were transformed to Counts Per Million (cpm) values and then normalized using the TMM normalization method from the edgeR package (Robinson, McCarthy et al., 2010) (version 3.2) was used. Normalized genes were then used as an input for the Voom transformation method implemented in the *limma* package [28] in R 3.4.4. Genes with low expression values (cpm < 2 in more than 50% of the samples) were excluded from differential gene expression analysis. A linear-model was fit and differentially expressed genes were calculated across all samples involved in ADPKD progression and treated vs. untreated ADPKD samples. Raw data was deposited in ArrayExpress and given the following identifier E-MTAB-8086.

Validation datasets [15,29] were acquired from GEO (ID: GSE72554 and GSE7869) and further processed using *limma* for the identification of the differentially expressed genes in each of the different mice groups. For the data of Menezes et al., we compared the resultant lists of differentially expressed genes with the different clusters involved in ADPKD progression using the representation factor. The representation factor is the number of overlapping genes divided by the expected number of overlapping genes drawn from two independent groups. A representation factor > 1 indicates more overlap than expected of two independent groups, a representation factor < 1 indicates less overlap than expected, and a representation factor of 1 indicates that the two groups have the same overlap for independent groups of genes. For the data of Song et al., we combined the differentially expressed genes (P-value < 0.05, t-statistics) of the small and medium cysts and processed them using the method detailed in "Annotation of Gene Expression Profiles" sub-section.

2.2.2. Principal component analysis (PCA)

Samples involved in ADPKD progression were selected and prepared for Principal Component Analysis (PCA). Briefly, the above noise level voom-transformed gene expression values were organized in a data matrix and given as an input for the *ir.pca* function in R. The loadings of the different principal components were plotted in a 2-dimensional plot using the *ggplot2* package in R.

2.2.3. Gene expression clustering

Hierarchical clustering was applied on all differentially expressed genes resulting from the pairwise comparisons of all samples involved in ADPKD progression (*FDR < 0.005,* Supplementary Table 1*).* The *hclust* package in R was applied on the euclidean distance matrix calculated using the *dist* R function. Utilizing the *cutree* function implemented in R, the resultant clustering tree was cut into 15 clusters. For each cluster, all gene members were plotted. In addition, the average gene expression pattern was based on the averaged expression values at each time-point.

2.2.4. Annotation of gene expression profiles

We annotated the resulting gene expression profiles using the GeneTrail2 v1.5 tool [30]. We ran the over-representation analysis against the Wikipathways database. We used all expressed genes above noise level as background and accepted enriched terms with *P-value* < 0.05.

2.2.5. Drug targets acquisition and prioritization

All high quality data on the selected protein targets were acquired from ChEMBL release 22 [31]. High quality was defined as follows: data points with a ChEMBL confidence score of 9 (direct single protein target assigned), with a pChEMBL activity value, and having >= 30 compound measurements per protein. The pChEMBL value is the negative logarithm of activity in molar for curve fitted activity values such as Ki, IC50, EC50, AC50, XC50. Furthermore, only human proteins were considered. This led to a total of 990 protein targets (directly assigned targets), and 356,396 interactions with 240,433 compounds. *Mus musculus* gene identifiers were converted to the homologous *homo sapiens* identifiers using the BioMart tool on the Ensembl website [32], and cross-checked with the *Homo sapiens* drug targets. We prioritized the resulting drug targets from the ChEMBL database [33] through several filtering steps based on a couple of criteria. First the overlap between this set and the PKD progression genes was kept, a total of 168 protein targets with 54,698 annotated bioactivities, through 48,050 small molecules. Subsequently only drug targets that were annotated to small molecules that have been tested in phases 2, 3 or 4 of clinical trials were kept. This was aimed at keeping molecules that have passed phase 1, which is aimed at determining if a drug is safe for efficacy testing in phases 2 and 3, phases 4 represents approved and marketed drugs. Secondly, we filtered targets that have antineoplastic activity based on the Anatomical Therapeutic Chemical (ATC) Classification System [https://www.who.int/classifications/atcddd]. Thirdly, for all remaining drugs and targets, we investigated for each drug, its mode of action in relation to each of its remaining targets and compared this to the direction of deregulation in the PKD Progression. When a drug has a conflicting mode of action to what is needed to correct the target's expression in ADPKD, that drug received low priority. For example, if drugA is an agonist to an up-regulated target in PKD, drugA would be excluded (or receive low priority). We kept the drugs that did not have a known mode of action. Fourthly, for the remaining targets, we gave the highest priority to drug targets that were dysregulated in the early phases of the disease, followed by moderate phases and finally advanced phases.

2.3. 3D cyst drug screening

The 3D cyst culture assay has been performed with *Pkd1*-KO mouse-inner medullary collecting duct (mIMCD3) cells (mIMRFNPKD 5E4) as described previously [34]. In short: mIMRFNPKD cells were mixed with Cyst-Gel (OcellO, Leiden, The Netherlands) to a final concentration of 150,000 cells/mL. 15μl of cell-gel mix was pipetted to 384-well plates (Greiner Clear, Greiner Bio-One B.V.) using a CyBio Felix 96/60 robotic liquid dispenser (Analyik Jena AG). After gel polymerization at 37 °C for 30 min, 33 μL culture medium was added to each well. Cells were grown in gel for 96 h, after which the cells were co-exposed with forskolin (Calbiochem) and one of the following molecules: Rapamycin (SelleckChem, S1039), Staurosporin (SelleckChem, S1421); Birinapant (Bioconnect, PK-CA577-2597−1), Gamma-Linolenic acid (Sanbio, 90,220−50), Eicosapentaenoic acid (Sanbio, 90,110−50), Meclofenamic Acid (Sanbio, 70,550−1), Zileuton (Sanbio, 10,006,967−10) and Indometacin (Sanbio, 70,270−1). Rapamycin and Staurosporin were used as cyst swelling inhibiting or toxic control respectively. All conditions were tested in quadruplicate. After 72 h, cultures were fixed with 4% Formaldehyde (Sigma Aldrich) and simultaneously permeabilized with 0.2% Triton-X100 (Sigma

Aldrich) and stained with 0.25 M rhodamine-phalloidin (Sigma Aldrich) and 0.1% Hoechst 33,258 (Sigma Aldrich) in 1x PBS (Sigma Aldrich) for 12 h at 4C, protected from light. Imaging was done using Molecular Devices ImageXpress Micro XLS (Molecular Devices) with a 4x NIKON objective. For each well, 30 images in the Z- direction (50 μm apart) were made for both channels. Each image captures the whole well area. Image analysis for actin and nuclei was performed using Ominer analysis software (OcellO BV.) integrated in KNIME Analytics Platform (Konstanz, Germany, http://www.knime.org/). Further data analysis was also done in KNIME. The main readout for efficacy, "cyst area", was calculated per well as the average of the area in px of each object in every in-focus plain. This measurement was then normalized to positive (100%) and negative control (0%). The parameters used for toxicity; "nuclei area" and "nuclei roundness" were calculated in a similar fashion, "fraction apoptotic nuclei" was calculated as the amount of nuclei without actin signal relative to the total amount of nuclei, both as count-measurements. Graphs were made in Graphpad 6 (GraphPad Software, La Jolla, CA).

3. Results

3.1. Gene expression patterns associated with disease severity

To study the different phases of ADPKD progression, we have inactivated *Pkd1* at postnatal day 38 or 90 (adult phase) in the kidneys and harvested these animals at different time points after gene inactivation, resulting in five groups of mice with different disease stages. In these mice the largest group of cysts originate from the proximal tubules but cyst are also formed in distal tubules and collecting ducts [35]. *Pkd1cko* animals sacrificed after 2, 3 or 6 weeks after gene inactivation represent very early disease states. *Pkd1cko* animals sacrificed at 11 and 12 weeks after gene inactivation represent a moderate state of the disease and *Pkd1cko* animals sacrificed at 15 weeks after gene inactivation represent advanced disease. The kidney weight to body weight ratios (2KW/BW) of the five groups concurred with increasing disease severity in these samples (Fig. 1A, Supplementary Table 1).

We carried out RNA sequencing of the different groups of mice (Supplementary Figure 1). RNA was extracted from the five *Pkd1cko* and wild-type (WT) groups and cDNA sequenced on the Illumina 2500 Hiseq platform. Applying principal component analysis (PCA) on the gene expression profiles of these samples and plotting the first components revealed that most of the variance between samples could be attributed to differences in disease severity (principal component-1 (pc1), explaining 28% of the total variance, Fig. 1B). Extracting the 20 most influential genes in component-1 and plotting their expression in all disease progressing samples showed that these genes strongly correlated with disease progression (average Spearman's rank correlation coefficient = 0.7; Fig. 1C). Components-2 and 3 explained 26.7% and 8.3% of the variance respectively, where component 3 may reflect the variation between different mice.

3.2. Expression patterns associated with ADPKD progression

To gain fine-grained insights into the different patterns of gene expression during disease progression, we applied hierarchical clustering on the 2731 differentially expressed genes ($FDR < 0.005$) discriminating the groups of mice in different states of disease progression. In the first round of clustering we grouped the 2731 differentially expressed genes based on their expression patterns across the different disease progression stages into 15 clusters. This resulted in 12 gene clusters with distinct and coherent expression profiles, ranging in size from 32 to 367 genes. Additionally, three clusters contained genes that showed an aberrant pattern in just one of the samples. These clusters (cluster 3, 11 and 14) were removed from further analysis (Fig. 2A, Supplementary Table 2). The remaining 12 clusters were characterized by their gene expression patterns. For example, cluster 1 shows up-regulation in the early pre-cystic phases of the disease, particularly at 2wk, 3wk and 6wks after gene inactivation. Cluster 4 on the other hand includes genes that are up-regulated in the moderate phase of the disease starting from 11 weeks of gene inactivation. Cluster 10 is an example of a cluster that contains genes that are down-regulated in the advanced stages of the disease, at 12−15 weeks after gene inactivation. As we are interested in the three distinct phases of the disease (i.e. early, moderate or advanced), we further grouped the 12 clusters into 3 groups, where each of the new groups represents one of the three distinct phases, with genes up- or down- regulated particularly in early (n = 5 clusters), moderate (n = 4 clusters) or late (n = 3 clusters) phases of disease (Fig. 2A).

3.3. The expression patterns can be replicated in an independent study

Menezes et al. recently published a study of a different *Pkd1* knockout mouse model for ADPKD [15]. They included mouse samples at different disease stages namely, pre-cystic, cystic and severely cystic. We tested the statistical enrichment, using the representation factor (RF) method, of the genes in each of our three disease-stage groups were compared to the genes that are differentially expressed in pre-cystic, cystic and severely-cystic male mice of Menezes et al. As expected, the early dysregulated group demonstrated the strongest overlap with the pre-cystic groups in Menezes et al. study (Fig. 2B). Likewise, the moderate stage group showed greater overlap in the cystic and severely cystic groups. Similar patterns were observed in the advanced gene group, which was most consistent with the cystic and severely cystic groups (Fig. 2B). The strong overlap observed across disease stages was more evident in the up-regulated clusters compared to the down-regulated clusters. Taken together, these results reflect strong reproducibility of the expression patterns in an independent study. Since 2KW/BW is an accepted measurement of ADPKD disease stage and progression, we correlated the expression values in the 12 distinct clusters with 2KW/BW. The spearman coefficient plotted in Fig. 2C showed strong correlation of moderate and advanced stage clusters with 2KW/BW, while the early phase clusters had a weak correlation with 2KW/BW. This is expected, because the early ADPKD samples have 2KW/BW similar to that of the wild types.

3.4. Biological functions and pathways associated with ADPKD progression

To understand the biological functions involved in ADPKD progression, we looked for the over-represented pathways in each of the three disease phases, early, moderate and advanced. For each disease phase, we combined the genes of the clusters that belonged to that phase and used GeneTrail2 v1.5 Wikipathways database to annotate them ($FDR < 0.05$). Terms enriched ($FDR < 0.05$) in any of the three disease phases are shown in Fig. 3A and provided as Supplementary Table 3. Hierarchical clustering was used to distinguish pathways that were specifically enriched in the early, moderate or late phases of the disease, and the pathways that were dysregulated across all phases (Fig. 3B). Interestingly, even at the pre-cystic phases we observed dysregulation in metabolism in the form of dysregulated TCA cycle and fatty acid biosynthesis, as well as Wnt signaling. Additionally, we observed dysregulation in G13 signaling pathway that is involved in cytoskeletal remodeling in cells and is essential for receptor tyrosine kinase-induced migration of fibroblast and endothelial cells. In the moderate and advanced phases of the disease, proliferation-related and inflammation-related pathways were dominant. The oxidative stress pathway, p53 and DNA mismatch-repair pathways were clearly visible in the advanced phase, along with alterations in metabolism. TNFα and chemokine signaling were active during all phases, from pre-cystic to advanced PKD. Using the work of Song et al. 2009 as a

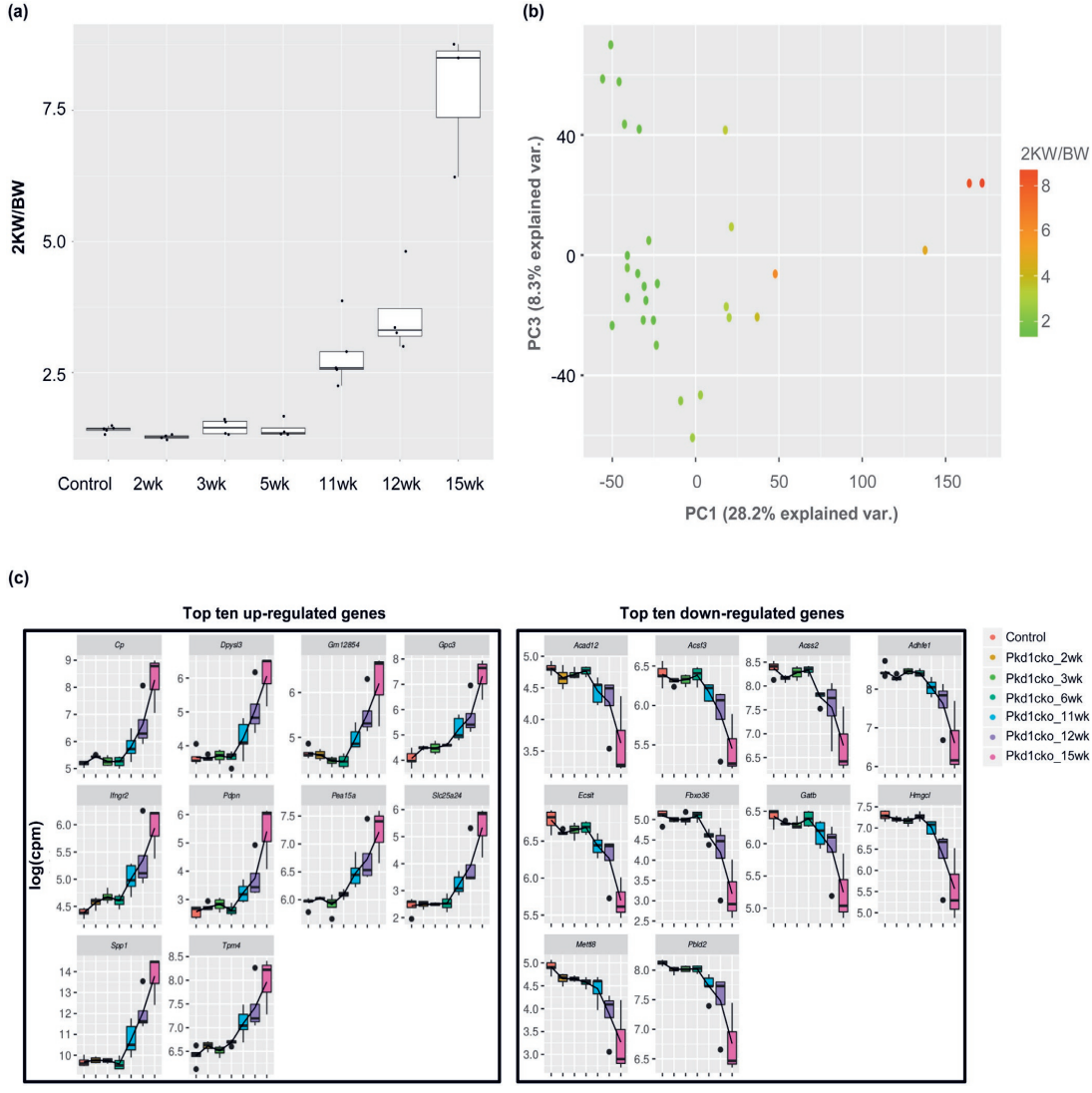

Fig. 1. Kidneys taken out at various disease stages show differences in expression profiles. (a) Boxplot representation of the 2KW/BW values for groups of *Pkd1cko* mice representing different phases of ADPKD with increasing disease severity. (b) Results from principal component analysis of the *Pkd1cko* samples. Shown are the loadings of plot of pc1 (x-axis) and pc3 (y-axis) of all samples. In the panel the samples are colored based on their 2KW/BW value. (c) Boxplots of the top 10 most up-regulated (left part) and the 10 most down-regulated genes (right part) during disease progression, as extracted from the loadings of the genes on pc1. Expression data are given as log2 (counts per million).

reference for human ADPKD, we confirmed the dysregulation of several of the aforementioned pathways in PKD patients. These include TCA cycle alterations, aberrant metabolism, active cytoskeleton remodeling and inflammation (Supplementary Table 3D).

3.5. Further selection of the ADPKD progression genes by evaluating response to therapy

We have previously shown that treating *Pkd1cko* P40 mice with Rapamycin and Curcumin and *Pkd1cko* P10 mice with soluble activin receptor IIB Fc (sActRIIB-Fc) significantly reduced kidney size and slowed the progression of ADPKD in mice [23,24,26] (Table 1). Here, we sequenced the RNA of the kidneys of these drug-treated mice using Illumina 2500 Hiseq platform and identified the differentially expressed genes (DEGs) between the treated and untreated samples. The curcumin treated samples are *Pkd1cko* P40 mice harvested at 11 weeks of age after gene inactivation; the same mouse model was treated with Rapamycin and harvested at two time-points, 12 weeks and 15 weeks after gene inactivation (Supplementary Figure 2). The soluble activin receptor-Fc fusion (sActRIIB-Fc) treatment was given to *Pkd1cko* P10 mice at two different time-points after tamoxifen treatment, starting at 0.3 weeks for the early-treated samples and at 2.1 weeks for the late treated mice (Supplementary Figure 2). Both groups were harvested at 4.1 weeks of age.

To balance the analysis between the different treatment groups, we took equally sized lists of the most differentially expressed genes (sorted on *P-value*). The size of the gene list was based on the treatment group with the lowest number DEGs (i.e. sActRIIB-Fc treatment), which is equal to 840 genes (P-value < 0.05 t-statistics, Table 1). For 1162 out of the 2731 genes that we identified to be

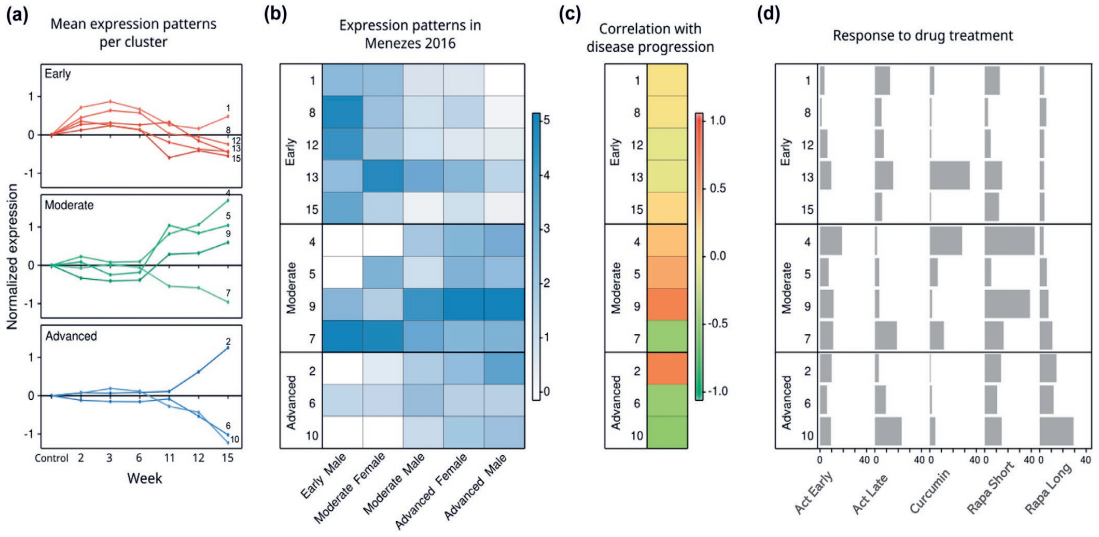

Fig. 2. 12 distinct expression patterns are associated with PKD progression. (a) The different expression patterns observed in *Pkd1cko* mice representing the progression of ADPKD towards end-stage renal disease (weeks after tamoxifen induction). For each cluster, mean log-transformed gene expression levels relative to the control mice that did not receive tamoxifen are plotted. The top panel represents the early dysregulated clusters, the middle panel represents the clusters dysregulated in the moderate to advanced stage and the bottom panel the clusters associated with the advanced stage of the disease. (b) Replication of expression profiles in an independent study. For each cluster, a representation factor reflecting the gene overlap of each cluster with the expression signatures from the five different mouse groups defined in the study by Menezes et al. [15] is given in a color representation. A representation score > 1 reflects enrichment. (c) Correlation of gene expression with disease progression. For each cluster, the average Spearman's correlation coefficient between the expression values of the genes in a cluster and the 2KW/BW ratio was calculated. Green represents a negative correlation while red reflects a positive correlation. Clusters that were dysregulated in an early stage have the lowest correlation with the 2KW/BW increase, suggesting they follow a different trend in disease progression. (d) Association of cluster with drug response. A bar chart representation for each cluster showing the proportion of the 2731 genes that were also affected by one of the drug treatments: sActRIIB-Fc early (Act Early) and late (Act Late), curcumin, rapamycin short (Rapa Short) and long (Rapa Long). The x-axis represents the % of genes that were significanltly dysregulated ($P < 0.05$) due to the drug treatments per cluster per drug treatment.

involved in ADPKD progression, the expression was normalized after at least one drug treatment, i.e. upregulated genes were not or less increased after treatment, or downregulated were not or less decreased after treatment (Supplementary Table 2). Since the drug treatments were effective in slowing disease progression, these genes reflect the healthier state of the kidneys upon drug treatment. By focusing on genes that respond to therapy we strengthen the involvement of the genes in disease progression and as potential target to

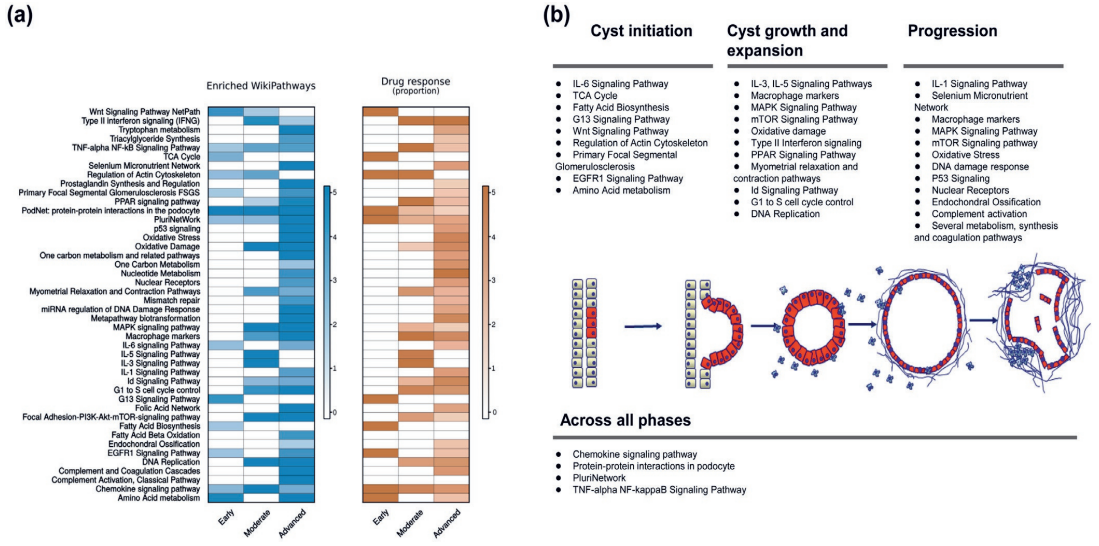

Fig. 3. Pathways associated with disease progression and drug response. (a) A heatmap representation of the molecular pathways significantly enriched in the different stages of PKD (left). For each cluster category from Fig. 2A, the significantly enriched Wikipathways were obtained (*FDR < 0.05)* and plotted in the heatmap. Color scale reflects the representation factor in the different phases of ADPKD (Early, Moderate and Advanced). On the right, a heatmap representation of the pathways that are enriched with significantly dysregulated genes after drug treatment. Enrichment was established based on the representation (RF) factor calculation, where pathways that had RF >= 1 are considered significant. (b) A schematic representation of the different pathways involved in ADPKD's progression. Pathways are selected based on their Wikipathways significance (*FDR)* across the different disease stages in part A.

Table 1
The results of the RNA-Sequencing results of the drug treated samples (Curcumin, Rapamycin and sActRIIB-Fc). Significant genes (P-value < 0.05, t-statistics) were identified based on the comparison of the drug treated samples to the non-treated control (see methods for details).

PubChem CID	Drug name and drug treatment (*Supplementary figure 2*)	No. of genes significant genes (*P-value < 0.05, t-statistics*)	Normalized no. of genes compared to PKD progression[a]	No. of genes found in PKD progression clusters
969,516	Curcumin	8030	840	503
5,284,616	Rapamycin Short	1600	840	441
5,284,616	Rapamycin Long	1250	840	322
NA	sActRIIB-Fc late-short treatment	840	840	270
NA	sActRIIB-Fc early-long treatment	4200	840	365

[a] Number is based on the lowest maximum of significant genes. This belongs to sActRIIB-Fc Late-Short treatment.

identify novel drugs to treat ADPKD (Figure-2D Fig. 2D). Table 1 summarizes the number of genes differentially expressed in the treated samples and involved in ADPKD progression.

3.6. Identifying drug targets from the genes associated with ADPKD progression

To identify candidate drugs that might have a favorable effect on ADPKD, we screened the ChEMBL database for drug-protein interactions. From ChEMBL, we only used high quality drug-protein target interactions (See Methods). This generated a list of 990 protein targets (directly assigned targets), and 356,396 interactions with 240,433 compounds. We compared these drug targets to our set of differentially expressed genes. From the total set of 1162 genes that were involved in ADPKD progression, 168 genes were annotated in ChEMBL as candidate drug targets and had enough high-quality bioactivity information to be used in our subsequent analysis (Fig. 4A). These 168 genes were targeted by 48,050 small molecules (Supplementary Table 4: Step 11).

As we were interested in the set of candidate drugs that can be repurposed for ADPKD, we extracted compounds that were tested in 2nd, 3rd or 4th phase clinical trials. 544 out of the 48,050 compounds met these criteria, and these compounds interacted with 111 of our selected targets (Supplementary Table 4: Step 12). Further restriction of the targets by including only those for which the mRNA levels were normalized by treatment with one or more of the three different drugs that slowed cyst formation in our preclinical models. This resulted in a set of 63 unique candidate drug targets interacting with 339 compounds (candidate drugs) (Supplementary Table 5A). For each candidate drug we have obtained its classification in the Anatomical Therapeutic Chemical (ATC) Classification System (https://www.who.int/medicines/regulation/medicines-safety/toolkit_atc/en/) and removed drugs with antineoplastic classification, because they may show too serious side effects during long-term treatment for PKD. 32 unique targets and 116 candidate drugs passed this filtering (Supplementary Table 5B). For each remaining drug-target interaction we looked for information on the mode of action (MoA) of the drug. However, this information was only available for a small subset of these interactions (12 targets and 23 drugs), and we filtered out the targets that had a conflicting direction of dysregulation in relation to the its drug MoA (3 targets were filtered out). We arrived at 29 genes that could serve as a target for drug repurposing in ADPKD (Table 2; Supplementary Table 5B).

3.7. Selection of candidate drugs

Analyzing the remaining 29 targets, we identified several that were previously linked to ADPKD treatment. For example, Suramin

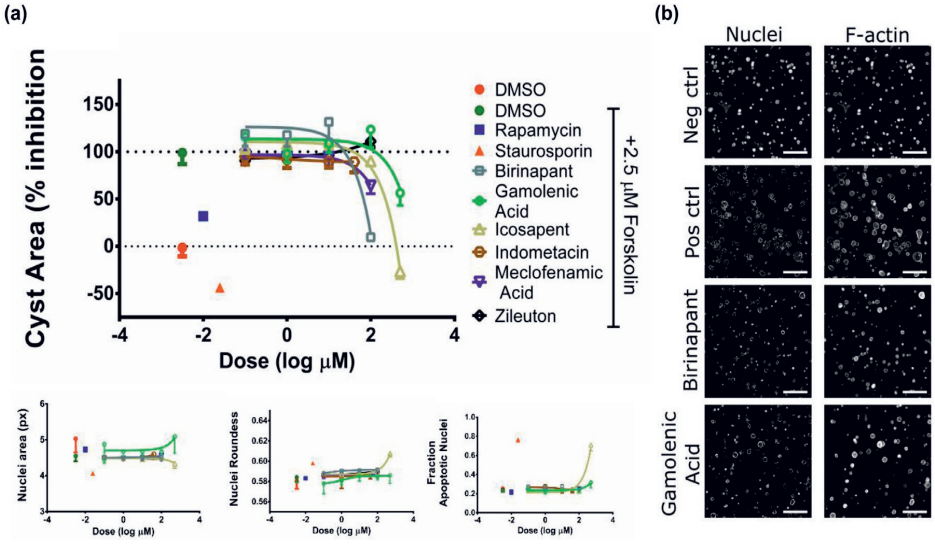

Fig. 4. 3D-cysts assay of candidate compounds. (a) Top: Quantification of cyst size of the tested compounds normalized to forskolin induced swelling. Reference compounds rapamycin (0.01 μM) and staurosporin (0.25 μM) reduce cyst size, as well as brinapant, Gamolenic Acid, icosapent and Meclofenamic Acid at highest tested concentration of 100 μM, 500 μM, 500 μM and 100 μM respectively (N = 4 wells). Bottom: Assessment of staurosporin-like induction of toxicity. Graphs representing average nuclei area, nuclei roundness and the fraction of nuclei that are apoptotic show changes for reference compound staurosporin and for icosapent. (b)Representative images of positive and negative control and two of the test compounds at highest tested dose; 100 μM for brinapant and 500 μM for Gamolenic Acid. Each scalebar is 400 μM.

Table 2
The 29 drug prioritized targets grouped based on their gene category according to MsigDB (*Supplementary Table 5B*).

Gene category	Count of genes	Genes
Protein kinases	3	CDK1 PRKCB PRKCZ
Transcription factors	3	PPARD STAT3 THRA
Cytokines and growth factors/receptors	2	CCL2 CCR2
Other	21	SLC1A1 PTGER3 AKR1B10 LGALS3 BIRC2 HSD17B2 P2RX7 PLD2 CYP2J2 MGLL FKBP4 PTGES ALOX5AP P2RY6 AKR1A1 MAPT CYP51A1 TRAP1 AKR1C1 AKR1C3 AKR1C2

Hexasodium a non-specific inhibitor of P2 receptors inhibiting P2Y and P2X receptors reduced cyst growth in a 3D cysts models while a P2 × 7 receptor antagonist as well as gene knock-down were previously shown to inhibit cystogenesis in a zebrafish model for polycystic kidney disease [36,37]. Suramin is also an antagonist of IL-6, known to inhibit renal fibrosis in chronic kidney disease in rats [38]. Another identified drug, the PPARγ agonist Rosiglitazone, was shown to be effective in animal models for PKD [39–41]. Pioglitazone, a close PPARγ agonist to Rosiglitazone, is currently undergoing clinical trials for ADPKD [ClinicalTrials.gov Identifier: NCT02697617]. Another prioritized candidate drug was the second mitochondrial-derived activator of caspases (SMAC)-mimetic GT13072, which slowed down PKD progression in two *Pkd1* mouse models [42]. Icosapent also known as ethyl eicosapentaenoic acid, another candidate drug prioritized by our analysis, was shown to reduce PKD severity in a mouse model, but this could not be confirmed in a small clinical trial [43]. Additionally, three candidate drugs, the spleen tyrosine kinase inhibitors entospletinib and R-406, and the polo-like kinase 1 inhibitor BI-2536 were previously shown to be effective in a 3D Cyst screen of the Selleckchem library of compounds [34]. Collectively, these previous findings support our approach in successfully identifying lead compounds for ADPKD drug repurposing.

To identify new candidate drugs for the ADPKD drug development pipeline, we evaluated six compounds in a 3D Cyst assay similar to that performed by Booij et al. [34]. The six candidate drugs, Zileuton, Indometacin, Meclofenamic Acid, Gamolenic Acid, icosapent and Birinapant (Table 3) were selected based on additional evidence for the potential therapeutic potential for ADPKD present in the Euretos Knowledge Platform (https://www.euretos.com/) and the scientific literature.

3.8. Wet-lab validation of selected candidate drugs

To test the selected drug candidates, we grew renal epithelial cells (mIMRFNPKD 5E4) in a 3D-gel matrix to allow cyst formation. After 96 h, cysts were co-exposed to forskolin, to induce cyst swelling, and the selected compounds for a period of 72 h. Rapamycin, shown before to reduce cyst swelling in several models [34,44], was used as a positive control for cyst swelling inhibition and demonstrated the expected reduction in cyst size (Fig. 4A). Of the selected drug candidates, Meclofenamic Acid, Gamolenic Acid, icosapent and Birinapant slowed cyst growth at the highest concentration tested; 100 μM, 500 μM, 500 μM and 100 μM respectively. Zileuton and indomethacin, however were not as effective (Fig. 4A, top), showing no effect on cyst size on any of the tested concentration up to 100 μM and 40 μM respectively. Birinapant was the most potent compound, with 50% inhibition of cyst swelling around 50 μM. These results were validated in an independent experiment (Supplementary Figure 4). To be able to distinguish true swelling inhibiting properties from severe toxicity, which also leads to reduced cyst size, staurosporin was included as a prototypic toxic compound at. Looking at the effect of staurosporin at 0.25 μM on phenotypic parameters such as nucleus size and shape as well as nucleus fractionation, there is clear induction of cytotoxicity. Of the selected compounds however, only icosapent shows similar kind of phenotypic changes, starting at a concentration of around 100 μM (Fig. 4A, bottom). Representative images of treatment effect can be found in Fig. 4B. These results indicate that 3 out of 6 novel compounds selected through our approach demonstrated to be able to inhibit cyst swelling in vitro without apparent toxicity.

4. Discussion

In this study we combined comprehensive gene expression profiling and bioinformatics, with cheminformatics to identify drugs for repurposing and targets to further explore for ADPKD treatment. Our approach is based on an innovative strategy that combines transcriptomics sequencing of different disease states of ADPKD and drug assays databases to arrive to a list of candidate drugs that could have a treatment potential for PKD. Our methodology zooms-in on a set of genes involved in ADPKD progression and proposes candidate drugs that could alter disease progression by targeting relevant genes. Our

Table 3
Drugs selected for validation in 3D Cyst experiment and their results.

PubChem CID	Drug name	Targets (pChEMBL value[a])	Results in 3D cyst assay	ATC code (Level 4)
49,836,020	Birinapant	BIRC2 (7.3)	Effective	n/a
5,280,933	Gamma-Linolenic acid (Gamolenic Acid)	PPARD (6.1)	Effective	D11AX
446,284	Eicosapentaenoic acid (Icosapent)	PPARD (5.4)	Effective	n/a
4037	Meclofenamic Acid	AKR1C3 (6.3), AKR1C1 (5.5), AKR1C2 (5.1)	Effective	M01AG
60,490	Zileuton	ALOX5AP (5.5)	Not effective	n/a
3715	Indometacin	AKR1C3 (6.2) , PTGES (4.4) , AKR1C2 (4.3)	Not effective	C01EB M01AB M02AA S01BC

[a] pChEMBL is a combination of a number of roughly comparable measures of half-maximal response concentration/potency/affinity to be compared on a negative logarithmic scale: -Log(molar IC50, XC50, EC50, AC50, Ki, Kd or Potency). We have tested compounds at 100 μM (pCHEMBL value 4). We have selected targets for which the affinity was a pCHEMBL value > 4.0.

work is of high relevance to PKD patients since they have limited treatment options. Tolvaptan (Jinarc), the only treatment now available, has limited efficacy, and side-effects like massive diuresis may limit patient adherence [20,45]. Therefore, there is a need for drugs that specifically target the formation and growth of cysts to slow down or halt disease progression. Given the complexity of altered signaling in cyst-lining epithelia, a broad range of potential targets are available, and drug-repurposing is a relative fast strategy for the development of new treatments.

We used a tamoxifen-induced *Pkd1cko* mouse model to generate expression profiles of the kidneys of 7 groups of mice with varying levels of disease progression. Using clustering techniques, we arrived at groups of genes that show altered expression in mild, moderate and advanced stages of the disease, each characterized by increased or reduced activation of certain pathways and pathogenic processes. In the early stage, the TCA cycle, fatty acid biosynthesis, EGFR signaling and G13 signaling were most significantly altered, indicating altered metabolism, proliferation and cytoskeletal remodeling, confirming previous studies in PKD [15,46]. In the moderate phase, we specifically observed increased MAPK and mTOR signaling, both involved in a broad range of cellular processes including cell proliferation and cell stress-related pathways (MAPK) or cell growth, proliferation, protein translation, autophagy, as well as actin cytoskeleton remodeling and apoptosis (mTOR) [23,47,48]. Additionally, at this stage we observed an up-regulation of cytokines such as IL-5 and IL-3, corresponding to inflammatory infiltrates and an active injury response. Inflammation and associated fibrosis became even more prominent in the advanced phase with increased expression of macrophage markers [49,50]. Furthermore, in the late-phase we see evidence of severe cell damage and tissue injury response with the up-regulation of pathways involved in oxidative stress, DNA damage response, and P53 signaling [29,51].

To arrive to a set of candidate drugs that could be repurposed for ADPKD, we took advantage of ChEMBL, where we identified molecules that target genes of the ADPKD progression profile. The advantage of using ChEMBL is that it is based on primary scientific literature, allowing us to validate the source of the bioactivity when needed. However, it should be noted that a similar approach could be envisioned with PubChem Bioassay or another source of biological activities. To make sure that the drug target relationships are of high quality we followed a series of filtering steps that led to 116 molecules binding to 29 genes. It is known that on average *approved* drugs show activity for 6 protein targets, so our selected molecules cannot be considered more promiscuous than normal in particular given that they have gone through phase 1 clinical trials [52]. Our filtering steps aim to minimize the number of 'wet-lab' experiments by focusing on only the most relevant and most confident information from literature. To be able to repurpose approved drugs, we did not only retrieve bioactivity data but also retrieved the primary (mode of action) target of each drug. Hence, we also included associated gene targets for approved drug that do not directly relate to the working mechanism described in the literature. As we included only drugs that are used in phases 2, 3 or 4 clinical trials and then filtered out drugs that have antineoplastic effects, we aimed to optimize our selection of drug repurposing candidates. The rationale being that compounds showing toxicity effects in phase 1 drugs known to kill (tumor) cells are less suitable for chronic administration to ADPKD patients. Out of the 116 candidate drugs that we prioritized for ADPKD treatment, we identified 5 molecules that were previously linked to PKD in 3D cultures and/or preclinical studies. More research is required to decide for further clinical development of these drugs/drug targets. Using a 3D-cyst drug screen assay, we have tested the effect of a further 6 drugs on cyst size at four or five dosages. In all cases the screening concentration we used was higher than the noted pChEMBL value (indicating that more than 50% of the compound was bound to the targets). 4 out of the 6 tested drugs had a positive impact on cyst size (decreased cyst size compared to controls). This became more evident at the high dosage, which might suggest a certain toxic effect on the cyst. We further analyzed the toxicological effects of these drugs and our initial toxicology analysis, revealed toxic effects of only 1 of the tested drugs.

The three remaining effective and nontoxic compounds are Meclofenamic Acid, Gamolenic Acid and Birinapant. From Table 3 it follows that the following targets could be responsible for the observed activity of these three compounds: BIRC2, PPARD, and AKR1C1. BIRC2 is the only known target for Birinapant and is in the identified targets. PPARD is a target for both Gamolenic Acid and Icosapent (and in the list of identified targets). AKR1C1, AKR1C2, and AKR1C3 are all in the list of identified targets and have an affinity for the active Meclofenamic Acid. However, the inactive compound Indometacin also has an affinity for AKR1C2 and AKR1C3, ruling them out as the prime targets for Meclofenamic Acid. Finally, PTGES and ALOX5AP seem not to be relevant targets as the inactive compounds Indometacin and Zileuton have affinity for them. It should be noted that the here retrieved targets represent only the targets for which activity was measured in the scientific literature; absence of these measurements does not demonstrate the absence of potential affinity. Moreover, the tested compounds may also have more targets on which they may demonstrate affinity (Supplementary Table 6). However, we selected in our approach only genes that were shown to be affected in ADPKD, which is not true for the other targets listed in Supplementary Table 6.

For the identified drugs we were also able to obtain more relevant information from literature, interestingly all these results are in line with our findings from Table 3. Meclofenamic Acid has been identified to target aldo-keto reductase family 1, which is implicated in steroid metabolism [53], which was reported to be involved in cyst development in cpk rat, a PKD model [54]. Gamolenic Acid has been selected based on PPARδ, which controls an array of metabolic genes involved in glucose homeostasis and fatty acid synthesis/storage, mobilization and catabolism. For other PPAR family members, PPARα and PPARγ, are being studied in (pre)clinical trials for PKD [40,55]. Birinapant is a SMAC mimetic and known modulator of apoptosis, which binds to and inhibits the activity of Inhibitors of Apoptosis Proteins (IAPs), including BIRC2(=cIAP1) thereby freeing caspases to activate apoptosis [56]. Another SMAC mimetic, GT13072, was previously shown to slow down PKD progression in *Pkd1* mouse models [42]. Overall, these drug candidates are relevant to the molecular events involved in ADPKD progression. However, further testing and pre-clinical experiments are needed to determine the efficacy of these drugs for ADPKD treatment.

To our knowledge this is the first drug repurposing effort in ADPKD at this scale. It expands on the previous transcriptomics efforts performed by others in the field. In this study we used deep RNA-sequencing of ADPKD transcriptomics across multiple disease stages, rather than microarrays [15,29,57,58]. The aforementioned studies differ in several elements, most notably their source of studied samples. Where we and Menezes et al. used adult *Pkd1* mutant mice, Pandey et al. used embryonic kidneys of *Pkd1* mutants and both Song et al. and de Almedia et al. used patient obtained ADPKD kidneys of ADPKD patients. Despite these differences, comparable dysregulated pathways have been reported. In all studies, abnormalities in metabolism, cell cycle and cell death are observed. Our results suggest that irregulates in metabolism and cell growth could play a role in early cyst development. Furthermore, we sequenced drug-induced ADPKD models to target progression involved genes at a higher precision, and thus enabling enhanced drug-repurposing. Our method screens thousands of approved drugs for their potential to treat ADPKD, expanding the work of others that focused on studying a selected number of drugs [59–63].

Although our approach is supported by wet-lab and *in silico* experiments, we acknowledge several limitations of our study. (1).

For the adult onset PKD mice, we only included males, while several results suggest ADPKD presentation differences between males and females [64]. Despite the differences in progression rates, gene network analyses revealed that the underlying mechanisms of PKD progression between male and female mice do not differ [15]; (2) Our starting point was gene expression data, while not all molecular processes act through changes in gene expression. Stage specific proteomics data and analysis of posttranslational modifications would be needed to obtain a more comprehensive insight in the molecular pathways associated with disease progression and would improve the quality of our drug predictions; (3) Drugs and their targets are biased towards the most studied drugs, diseases, and proteins (i.e. enzymes and G protein coupled receptors make up more than 75% of the data), while less-well characterized drugs may constitute equally good candidates for drug repurposing strategies [65]; (4) Further functional wet-lab experiments would be needed to determine the exact contribution of each gene to ADPKD progression and cyst growth. As more data will be implemented in ChEMBL and other biomedical database in the future, the power of this approach will increase. In addition, this approach is widely applicable to other diseases as well, provided that large scale high quality transcriptomic/proteomics data is available to be compared to databases cataloging drug affinity and activity towards a broad range of protein targets.

Funding sources

The research leading to these results has received funding from the People Program (Marie Curie Actions) of the European Union's Seventh Framework Program FP7/20072013 under Research Executive Agency Grant Agreement 317246 and under grant agreement 305444 'RD−Connect', and the Dutch Technology Foundation Stichting Technische Wetenschappen Project 11823, which is part of The Netherlands Organization for Scientific Research and a grant from the Dutch Kidney Foundation (17PhD02). The funders didn't have any role in study design, data collection, data analysis, interpretation, writing of the report.

Author contributions

T.B.M., W.N.L, P.A.C.t.H. and D.J.M.P. conceived and designed research; T.B.M., W.N.L., H.B., Z.G. and G.J.P.V.W. performed experiments; T.B.M., W.N.L., H.B., Z.G., G.J.P.V.W. and K.M.H. analyzed data; T.B.M., W.N.L., L.S.P., P.A.C.t.H. and D.J.M.P. interpreted results of experiments; T.B.M. and H.B. prepared figures; T.B.M. drafted manuscript; T.B.M., W.N.L., H.B. Z.G, K.M.H., G.J.P.V.W., L.S.P. P.A.C.t.H. and D.J.M.P. edited and revised manuscript; T.B.M., W.N.L., Z.G., H.B., K.M.H., G.J.P.V.W., L.S.P., P.A.C.t.H. and D.J.M.P. approved final version of manuscript.

Declaration of Competing Interest

Kristina M. Hettne performed paid consultancy between November 1, 2015 and March 31, 2018 for Euretos B.V, a startup founded in 2012 that develops knowledge management and discovery services for the life sciences, with the Euretos Knowledge Platform as a marketed product. Leo Price is a founder shareholder and Hester Bange employee at OcellO B.V., which operates in the PKD drug discovery field. All other authors have nothing to disclose.

Acknowledgements

We thank Mohammed Charrout for his contribution to the design of the figures.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ebiom.2019.11.046.

References

[1] Nosengo N. Can you teach old drugs new tricks? Nature 2016;534(7607):314–6.
[2] Dudley JT, Deshpande T, Butte AJ. Exploiting drug-disease relationships for computational drug repositioning. Brief Bioinform 2011;12(4):303–11.
[3] Weinstein JN, Myers TG, O'Connor PM, Friend SH, Fornace Jr AJ, Kohn KW, et al. An information-intensive approach to the molecular pharmacology of cancer. Science 1997;275(5298):343–9.
[4] Hughes T, Andrews B, Boone C. Old drugs, new tricks: using genetically sensitized yeast to reveal drug targets. Cell 2004;116(1):5–7.
[5] Booij TH, Leonhard WN, Bange H, Yan K, Fokkelman M, Plugge AJ, et al. In vitro 3d phenotypic drug screen identifies celastrol as an effective in vivo inhibitor of polycystic kidney disease. J Mol Cell Biol 2019.
[6] Mullen J, Cockell SJ, Woollard P, Wipat A. An integrated data driven approach to drug repositioning using gene-disease associations. PLoS ONE 2016;11(5):e0155811.
[7] Keiser MJ, Setola V, Irwin JJ, Laggner C, Abbas AI, Hufeisen SJ, et al. Predicting new molecular targets for known drugs. Nature 2009;462(7270):175–81.
[8] Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, et al. The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. Science 2006;313(5795):1929–35.
[9] Readhead B, Hartley BJ, Eastwood BJ, Collier DA, Evans D, Farias R, et al. Expression-based drug screening of neural progenitor cells from individuals with schizophrenia. Nat Commun 2018;9(1):4412.
[10] Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. Cell 2017;171(6) 1437-52 e17.
[11] Lanktree MB, Haghighi A, Guiard E, Iliuta IA, Song X, Harris PC, et al. Prevalence estimates of polycystic kidney and liver disease by population sequencing. J Am Soc Nephrol 2018;29(10):2593–600.
[12] Willey CJ, Blais JD, Hall AK, Krasa HB, Makin AJ, Czerwiec FS. Prevalence of autosomal dominant polycystic kidney disease in the European Union. Nephrol Dial Transplant 2017;32(8):1356–63.
[13] Audrezet MP, Cornec-Le Gall E, Chen JM, Redon S, Quere I, Creff J, et al. Autosomal dominant polycystic kidney disease: comprehensive mutation analysis of PKD1 and PKD2 in 700 unrelated patients. Hum Mutat 2012;33(8):1239–50.
[14] Spithoven EM, Kramer A, Meijer E, Orskov B, Wanner C, Abad JM, et al. 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 2014;29(Suppl 4) iv15-25.
[15] Menezes LF, Lin CC, Zhou F, Germino GG. Fatty acid oxidation is impaired in an orthologous mouse model of autosomal dominant polycystic kidney disease. EBioMedicine 2016;5:183–92.
[16] Harris PC, Torres VE. Genetic mechanisms and signaling pathways in autosomal dominant polycystic kidney disease. J Clin Invest 2014;124(6):2315–24.
[17] Serra AL, Poster D, Kistler AD, Krauer F, Raina S, Young J, et al. Sirolimus and kidney growth in autosomal dominant polycystic kidney disease. N Engl J Med 2010;363(9):820–9.
[18] Walz G, Budde K, Mannaa M, Nurnberger J, Wanner C, Sommerer C, et al. Everolimus in patients with autosomal dominant polycystic kidney disease. N Engl J Med 2010;363(9):830–40.
[19] Caroli A, Perico N, Perna A, Antiga L, Brambilla P, Pisani A, et al. Effect of longacting somatostatin analogue on kidney and cyst growth in autosomal dominant polycystic kidney disease (ALADIN): a randomised, placebo-controlled, multicentre trial. Lancet 2013;382(9903):1485–95.
[20] Torres VE, Chapman AB, Devuyst O, Gansevoort RT, Grantham JJ, Higashihara E, et al. Tolvaptan in patients with autosomal dominant polycystic kidney disease. N Engl J Med 2012;367(25):2407–18.
[21] Verbist B, Klambauer G, Vervoort L, Talloen W, Consortium Q, Shkedy Z, et al. Using transcriptomics to guide lead optimization in drug discovery projects: lessons learned from the QSTAR project. Drug Discov Today 2015;20(5):505–13.
[22] Dixit AB, Banerjee J, Srivastava A, Tripathi M, Sarkar C, Kakkar A, et al. RNA-seq analysis of hippocampal tissues reveals novel candidate genes for drug refractory epilepsy in patients with MTLE-HS. Genomics 2016;107(5):178–88.
[23] Novalic Z, van der Wal AM, Leonhard WN, Koehl G, Breuning MH, Geissler EK, et al. Dose-dependent effects of sirolimus on mTOR signaling and polycystic kidney disease. J Am Soc Nephrol 2012;23(5):842–53.
[24] Leonhard WN, Kunnen SJ, Plugge AJ, Pasternack A, Jianu SB, Veraar K, et al. Inhibition of activin signaling slows progression of polycystic kidney disease. J Am Soc Nephrol 2016;27(12):3589–99.
[25] Happe H, Leonhard WN, van der Wal A, van de Water B, Lantinga-van Leeuwen IS, Breuning MH, et al. Toxic tubular injury in kidneys from Pkd1-deletion mice accelerates cystogenesis accompanied by dysregulated planar cell polarity and canonical Wnt signaling pathways. Hum Mol Genet 2009;18(14):2532–42.
[26] Leonhard WN, van der Wal A, Novalic Z, Kunnen SJ, Gansevoort RT, Breuning MH, et al. Curcumin inhibits cystogenesis by simultaneous interference of multiple signaling pathways: in vivo evidence from a Pkd1-deletion model. Am J Physiol Renal Physiol 2011;300(5):F1193–202.

[27] Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 2013;14(4):R36.
[28] Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 2015;43(7):e47.
[29] Song X, Di Giovanni V, He N, Wang K, Ingram A, Rosenblum ND, et al. Systems biology of autosomal dominant polycystic kidney disease (ADPKD): computational identification of gene expression pathways and integrated regulatory networks. Hum Mol Genet 2009;18(13):2328–43.
[30] Backes C, Keller A, Kuentzer J, Kneissl B, Comtesse N, Elnakady YA, et al. GeneTrail–advanced gene set enrichment analysis. Nucleic Acids Res 2007;35(Web Server issue):W186–92.
[31] Papadatos G, Gaulton A, Hersey A, Overington JP. Activity, assay and target data curation and quality in the chembl database. J Comput Aided Mol Des 2015;29(9):885–96.
[32] Kersey PJ, Lawson D, Birney E, Derwent PS, Haimel M, Herrero J, et al. Ensembl genomes: extending Ensembl across the taxonomic space. Nucleic Acids Res 2010;38(Database issue):D563–9.
[33] Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, et al. ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res 2012;40(Database issue):D1100–7.
[34] Booij TH, Bange H, Leonhard WN, Yan K, Fokkelman M, Kunnen SJ, et al. High-Throughput phenotypic screening of kinase inhibitors to identify drug targets for polycystic kidney disease. SLAS Discov 2017;22(8):974–84.
[35] Leonhard WN, Happe H, Peters DJ. Variable Cyst development in autosomal dominant polycystic kidney disease: the biologic context. J Am Soc Nephrol 2016;27(12):3530–8.
[36] Buchholz B, Teschemacher B, Schley G, Schillers H, Eckardt KU. Formation of cysts by principal-like MDCK cells depends on the synergy of cAMP- and ATP-mediated fluid secretion. J Mol Med (Berl) 2011;89(3):251–61.
[37] Chang MY, Lu JK, Tian YC, Chen YC, Hung CC, Huang YH, et al. Inhibition of the P2 × 7 receptor reduces cystogenesis in PKD. J Am Soc Nephrol 2011;22(9):1696–706.
[38] Liu N, Tolbert E, Pang M, Ponnusamy M, Yan H, Zhuang S. Suramin inhibits renal fibrosis in chronic kidney disease. J Am Soc Nephrol 2011;22(6):1064–75.
[39] Flaig SM, Gattone VH, Blazer-Yost BL. Inhibition of cyst growth in PCK and Wpk rat models of polycystic kidney disease with low doses of peroxisome proliferator-activated receptor gamma agonists. J Transl Int Med 2016;4(3):118–26.
[40] Blazer-Yost BL, Haydon J, Eggleston-Gulyas T, Chen JH, Wang X, Gattone V, et al. Pioglitazone attenuates cystic burden in the PCK rodent model of polycystic kidney disease. PPAR Res 2010;2010:274376.
[41] Dai B, Liu Y, Mei C, Fu L, Xiong X, Zhang Y, et al. Rosiglitazone attenuates development of polycystic kidney disease and prolongs survival in Han:SPRD rats. Clin Sci (Lond) 2010;119(8):323–33.
[42] Fan LX, Zhou X, Sweeney Jr. WE, Wallace DP, Avner ED, Grantham JJ, et al. Smac-mimetic-induced epithelial cell death reduces the growth of renal cysts. J Am Soc Nephrol 2013;24(12):2010–22.
[43] Higashihara E, Nutahara K, Horie S, Muto S, Hosoya T, Hanaoka K, et al. The effect of eicosapentaenoic acid on renal function and volume in patients with ADPKD. Nephrol Dial Transplant 2008;23(9):2847–52.
[44] Tao Y, Kim J, Schrier RW, Edelstein CL. Rapamycin markedly slows disease progression in a rat model of polycystic kidney disease. J Am Soc Nephrol 2005;16(1):46–51.
[45] Torres VE, Chapman AB, Devuyst O, Gansevoort RT, Perrone RD, Koch G, et al. Tolvaptan in later-stage autosomal dominant polycystic kidney disease. N Engl J Med 2017;377(20):1930–42.
[46] Rowe I, Chiaravalli M, Mannella V, Ulisse V, Quilici G, Pema M, et al. Defective glucose metabolism in polycystic kidney disease identifies a new therapeutic strategy. Nat Med 2013;19(4):488–93.
[47] Yamaguchi T, Nagao S, Wallace DP, Belibi FA, Cowley BD, Pelling JC, et al. Cyclic AMP activates B-Raf and ERK in cyst epithelial cells from autosomal-dominant polycystic kidneys. Kidney Int 2003;63(6):1983–94.
[48] Shillingford JM, Murcia NS, Larson CH, Low SH, Hedgepeth R, Brown N, et al. The mTOR pathway is regulated by polycystin-1, and its inhibition reverses renal cystogenesis in polycystic kidney disease. Proc Natl Acad Sci U S A 2006;103(14):5466–71.
[49] Cassini MF, Kakade VR, Kurtz E, Sulkowski P, Glazer P, Torres R, et al. Mcp1 promotes macrophage-dependent cyst expansion in autosomal dominant polycystic kidney disease. J Am Soc Nephrol 2018;29(10):2471–81.
[50] Yang Y, Chen M, Zhou J, Lv J, Song S, Fu L, et al. Interactions between macrophages and cyst-lining epithelial cells promote kidney cyst growth in Pkd1-deficient mice. J Am Soc Nephrol 2018;29(9):2310–25.
[51] Malas TB, Formica C, Leonhard WN, Rao P, Granchi Z, Roos M, et al. Meta-analysis of polycystic kidney disease expression profiles defines strong involvement of injury repair processes. Am J Physiol Renal Physiol 2017;312(4):F806–F17.
[52] Hu Y, Bajorath J. High-resolution view of compound promiscuity. F1000Res 2013;2:144.
[53] Flanagan JU, Yosaatmadja Y, Teague RM, Chai MZ, Turnbull AP, Squire CJ. Crystal structures of three classes of non-steroidal anti-inflammatory drugs in complex with aldo-keto reductase 1C3. PLoS ONE 2012;7(8):e43965.
[54] Aziz N, Maxwell MM, Brenner BM. Coordinate regulation of 11 beta-HSD and Ke 6 genes in cpk mouse: implications for steroid metabolic defect in PKD. Am J Physiol 1994;267(5 Pt 2):F791–7.
[55] Lakhia R, Yheskel M, Flaten A, Quittner-Strom EB, Holland WL, Patel V. PPARalpha agonist fenofibrate enhances fatty acid beta-oxidation and attenuates polycystic kidney and liver disease in mice. Am J Physiol Renal Physiol 2018;314(1):F122–F31.
[56] Srivastava AK, Jaganathan S, Stephen L, Hollingshead MG, Layhee A, Damour E, et al. Effect of a smac mimetic (TL32711, birinapant) on the apoptotic program and apoptosis biomarkers examined with validated multiplex immunoassays fit for clinical use. Clin Cancer Res 2016;22(4):1000–10.
[57] de Almeida RM, Clendenon SG, Richards WG, Boedigheimer M, Damore M, Rossetti S, et al. Transcriptome analysis reveals manifold mechanisms of cyst development in ADPKD. Hum Genomics 2016;10(1):37.
[58] Pandey P, Qin S, Ho J, Zhou J, Kreidberg JA. Systems biology approach to identify transcriptome reprogramming and candidate microRNA targets during the progression of polycystic kidney disease. BMC Syst Biol 2011;5:56.
[59] Tesar V, Ciechanowski K, Pei Y, Barash I, Shannon M, Li R, et al. Bosutinib versus placebo for autosomal dominant polycystic kidney disease. J Am Soc Nephrol 2017;28(11):3404–13.
[60] Zhou X, Fan LX, Peters DJ, Trudel M, Bradner JE, Li X. Therapeutic targeting of BET bromodomain protein, Brd4, delays cyst growth in ADPKD. Hum Mol Genet 2015;24(14):3982–93.
[61] Weimbs T, Shillingford JM, Torres J, Kruger SL, Bourgeois BC. Emerging targeted strategies for the treatment of autosomal dominant polycystic kidney disease. Clin Kidney J 2018;11(Suppl 1):i27–38.
[62] Seeger-Nukpezah T, Proia DA, Egleston BL, Nikonova AS, Kent T, Cai KQ, et al. Inhibiting the HSP90 chaperone slows cyst growth in a mouse model of autosomal dominant polycystic kidney disease. Proc Natl Acad Sci U S A 2013;110(31):12786–91.
[63] Shillingford JM, Piontek KB, Germino GG, Weimbs T. Rapamycin ameliorates PKD resulting from conditional inactivation of Pkd1. J Am Soc Nephrol 2010;21(3):489–97.
[64] Hwang YH, Conklin J, Chan W, Roslin NM, Liu J, He N, et al. Refining genotype-phenotype correlation in autosomal dominant polycystic kidney disease. J Am Soc Nephrol 2016;27(6):1861–8.
[65] Lenselink EB, Ten Dijke N, Bongers B, Papadatos G, van Vlijmen HWT, Kowalczyk W, et al. Beyond the hype: deep neural networks outperform established methods using a ChEMBL bioactivity benchmark set. J Cheminform 2017;9(1):45.

CHAPTER 5
DRUG REPURPOSING USING A SEMANTIC KNOWLEDGE GRAPH

Tareq B. Malas , Roman Kudrin, Sergei Starikov, Peter A.C. ´'t Hoen, Dorien J.M. Peters, Marco Roos, Kristina M. Hettne

SWAT4LS, Tech. Rep.

December 2017

Drug Repurposing Using a Semantic Knowledge Graph

Tareq B. Malas[1], Roman Kudrin[1,2], Sergei Starikov[1,2], Peter A.C.´t Hoen[1], Dorien J.M. Peters[1], Marco Roos[1], Kristina M. Hettne[1]

[1] Department of Human Genetics, Leiden University Medical Center, 2300 RC Leiden, The Netherlands
[2] Faculty of Bioengineering and Bioinformatics, Moscow State University, 119234 Moscow, Russia

Abstract. Given the significant time and financial costs of developing a commercial drug, it remains important to constantly reform the drug discovery pipeline with novel technologies that can narrow the candidates down to the most promising lead compounds for clinical testing. Computational approaches are used to expedite the drug discovery processes. Semantic knowledge graphs can assist these computational approaches, because they connect different biological databases and reflect the relationships between genes, pathways and diseases. Here, we took advantage of the Euretos Knowledge Platform (EKP), a commercial database that integrates more than 170 different biological resources including DrugBank, and evaluated the usefulness of the underlying semantic knowledge graphs to predict novel drug-disease associations. As a positive set, we extracted 403 drug-disease associations from an independent resource, Metab2MeSH. An equally sized negative set was created by reshuffling of these drug-disease triples. For prediction of new associations, it is important to assess which paths in the knowledge graph optimally represent novel drug-disease associations. All 403 drug-disease associations in the positive and negative dataset were connected by at least one intermediate concept from 12 out of the 14 semantic types available in EKP. 'Chemicals & Drugs' was the most informative intermediate semantic type and distinguished the positive and negative sets best (Kolmogorov-Smirnov p-value: 7.4. 10-23). Also graph network features such as the total number of intermediate concepts (count), the number of different semantic categories (diversity), and the predicates connecting a drug-disease pair were successful in separating the positive from the negative sets. These features can be used to build a classifier for the prediction of novel drug-disease associations from the Euretos Knowledge Platform facilitating drug repurposing in preclinical research.

Keywords: Drug repurposing, drug discovery, Semantic graphs, network mining, machine learning.

Introduction

In silico methodologies are becoming more important in the modern-day drug discovery pipeline. Computational drug discovery techniques accelerated the identification of drug targets and significantly contributed to the different stages of drug development [1]. Most efforts are concentrated into developing methods for the prediction of drug target interactions that mitigate the expensive costs of experimental drug development and optimization [2]. Moreover, these methods are allowing for drug repurposing efforts that identify new therapeutic applications for existing drugs and reduce research cost and time due to the existing extensive clinical studies [2, 3].
Given that the majority of diseases cannot be explained by single-gene defects but by the coordinated functions of their complex gene networks, drug development needs to shift its

attention towards understanding network-based perspectives of disease mechanisms. Network-based approaches are providing important insights into the relationship between drugs and diseases. An investigation into the interaction between drug targets and disease genes revealed that they are not closely related [4]. Additionally, network-based approaches are showing promise in predicting novel targets and new uses for existing drugs [5]. Current network-based approaches rely on drug target profile similarities. These similarities are defined by either the number of targets two drugs share or the shortest paths between their interactomes. However, these studies focus only on using a limited number of databases related to protein drug targets, leaving a large amount of rich data untapped.

Semantic and text-mining approaches that screen hundreds of thousands of published literature articles have demonstrated the possibility of extracting concepts of biological meaning of various types. Semantic knowledge graphs are constructed to connect concepts of various types based utilizing a number of resources such as literature knowledge and biological databases. Such knowledge graphs can then be used to infer novel connections based on network mining methods [6, 7]. In addition to semantic connections, large efforts were made to integrate biological databases across gene, protein, pathway, disease and drug domains. The Euretos Knowledge Platform (EKP, http://www.euretos.com/) is a commercial database that integrates more than 170 different biological resources including semantic data [http://www.euretos.com/files/-EKPSources2017.pdf]. These data sources are used by EKP to build a large network of connected biological concepts. Disease and drug concepts in EKP are directly or indirectly connected based on prior knowledge found in publications and/or other databases. We expect that leveraging a large set of databases will enhance our drug discovery ability and avoid relying on a single source of information to associate drugs to diseases. Each semantic type provides us with an additional layer of information that can be exploited to identify novel drug disease associations.

In this work, we have taken advantage of the EKP to evaluate the usefulness of the underlying semantic knowledge graphs to predict novel drug-disease associations. With the current exponential growth in biological data, semantic knowledge graphs have a great potential for drug discovery.

Materials and Methods

Data Acquisition and Mapping in EKP

Drug disease pairs were acquired from Guney et al [8]. We specifically acquired the drug disease associations based on their analysis. We had 403 pairs of 239 drugs and 78 diseases that formed our positive “gold-standard” (GD) data. Randomly shuffling the positive dataset, we created 20 negative random datasets and used their average in the downstream analysis.

In EKP we first mapped the DrugBank IDs of the drugs in our datasets to drug concepts in EKP. We used full disease names to map the diseases in our dataset to disease concepts in EKP. Triples of drug disease pairs were identified in EKP if they were directly connected by at least one of the resources used in EKP (Figure-1). Predicates of drug

disease triples were classified as “relevant” if they belonged to one of the following categories: “treats”, “affects”, “prevents”, “disrupts”. The LUMC has a local installation of this knowledge graph for research purposes

Network Features

Network features were calculated for the intermediate concepts connecting drug disease pair. To evaluate if we could use the indirect associations to predict novel associations between drugs and diseases, we used the positive and negative datasets as follows. For each indirect association, we calculated a number of features and tested if these features could separate the two datasets. These features were calculated for each semantic subcategory (SubSemantic) available in EKP.

I. Count_normalized referred to as count in the following text:

$$(\text{“SubSemantic\_typeY”}) = X \div (y \times z) \quad (1)$$

X = total number of SubSemantic_typeY connecting the drug (y number of unique drugs making one drug concept) with disease (z number of unique diseases making one disease concept)

II. Diversity = The total number of unique SubSemantic categories connecting the drug and disease concepts per semantic type.

III. Predicates from the drug concept to the intermediate concept and from the intermediate concept to the disease concept were combined and referred to as “predicate path”. We used the Chi-square test to identify, within each semantic subcategory, the most enriched paths in the GD vs the negative dataset (cutoff p-value < 0.05). We filtered out paths that made up less than 1% of the total amount of paths within each semantic subcategory.

For I and II we used the Kolmogorov–Smirnov to test the similarity of the distribution of scores between the positive and the negative datasets (cutoff p-value < 0.05)

Results and Discussion

Concept Mapping and Direct Associations

We acquired the dataset of curated drug disease relationships (drugs used in the treatment of certain diseases) from Guney et al [8]. The GD dataset included 239 drugs, 78 diseases and 403 drug-disease pairs. For the negative dataset, we reshuffled the GD into 20 random datasets. The results of the negative datasets were averaged and compared to the GD.

We used DrugBank IDs available in the GD dataset to map drugs from the GD and negative datasets into EKP concepts and we used the full disease name to map diseases, since no unique identifier was supplied in the GD dataset. Out of 239 drugs,

235 were mapped successfully. All diseases were mapped successfully into EKP. Whendisease or drug term mapped to more than one concept in the EKP, this was corrected for (Figure-1).

Using the EKP we retrieved the triples for drug-disease pairs found in the GD and negative datasets. Each semantic triple consists of a subject-predicate-object, where the subject and the object refer to the drug and the disease respectively, and the predicate refers to the relationship connecting them. From the pairs found in the GD, 83% mapped to a triple in the EKP, whereas in the negative datasets 22% of the pairs mapped to a triple in the EKP. Moreover, from the mapped triples in the GD, 90% had a predicate type that we consider positive for a drug-disease association i.e. 'treats', compared to 75% in the negative datasets. These results demonstrate that the drug disease pairs in the GD and the negative datasets are different in two main aspects. 1). Most of the GD drug disease pairs could be represented in direct triples owing to prior knowledge of the pair's relationship. 2). The type of the predicates is different when comparing the triples of the GD and negative datasets, where the GD contains a higher proportion of the "relevant" predicates.

Evaluating the Indirect Drug-Disease Associations

As we are interested in drug repurposing, we were looking for novel associations between drugs and diseases. We utilized the indirect drug disease associations as a basis for our method, where we aim to mine the full EKP graph of indirect drug disease associations for strong candidates using network based features. To identify which features are useful, we used the GD and the negative datasets and evaluated several network features on the indirect associations retrieved from them. In the EKP, 14 semantic types are defined based on the semantic groups as defined by the Unified Medical Language System [9], with a number of semantic subcategories under each semantic type. Our analysis of indirect associations, i.e. drugs and diseases that connected via a third concept, was done per subsemantic category.

All 403 drug-disease associations in the GD and negative dataset were connected by at least one intermediate concept from the semantic types available in EKP. Out of the 14 possible semantic categories, 12 were found to connect a drug and a disease. We next evaluated which semantic and semantic subcategories were the most informative. Using the count diversity feature, defined as the total number of a certain intermediate concept connecting a drug disease pair, the semantic type 'Chemicals & Drugs' was the most informative intermediate semantic type and distinguished the positive and negative sets best (Kolmogorov-Smirnov p-value: $7.4 \cdot 10^{-23}$). Density plots of the count values per semantic and semantic subcategory in both the GD and the negative data reveal visually that the GD contained a higher number of indirect concepts in most semantic categories compared to the negative dataset, such as "Chemicals & Drugs", "Anatomy", "Disorders" and "Procedures" semantic categories (Figure-2A, Table-1).

Another feature we investigated was the diversity of the different semantic types connecting a drug disease pair. In this analysis we compared the total number of unique semantic categories and semantic subcategories in the drug disease pairs of the GD and

negative datasets. As observed for the count feature, the GD drug disease pairs displayed a higher semantic diversity in their intermediate concepts (Figure-2B).

We also investigated the predicate types that connect the indirect concept with the drug disease pairs. In this analysis we used two predicates, the one connecting the drug with the intermediate concept and the one connecting the intermediate concept with the disease concept. The combination of these two predicates in this order is referred to as the predicate path. Using the chi-squares test we investigated if there were predicate paths that are enriched in the GD and negative datasets. We found the most enriched paths in the “Amino Acid, Peptide or Protein” and “Pharmacologic Substance” semantic subcategories (Figure-2C). For example,, the path “drug □*is compared with* □ Pharmacologic Substance □ treats □Disease” that belongs to the “Pharmacologic Substance” semantic subcategory is strongly enriched in the GD that can be interpreted as drugs that are known to be similar in function or chemical properties can be repurposed for the same disease.

These results indicate that the type of, count and the predicates relating to the intermediate concepts connecting a drug and a disease pair were informative in differentiating positive and negative datasets. The added values of using a diverse set of semantic categories was demonstrated. In the count feature, we found almost all semantic categories shifted towards higher values in the GD when compared to the negative data. Additionally, the diversity feature revealed that the GD tends to have a higher number of semantic categories and subcategories as intermediate concepts connecting drugs and diseases. Having the ‘Chemicals & Drugs’ as the most differentiating semantic category also demonstrates the importance of looking at drug properties and not completely relying on the drug targets.

In contrast to other tools, our methodology is different in a number of ways. The quantity and diversity of databases that we included is larger and the content much richer than other comparable tools. In terms of quantity we have taken advantage of EKP that integrates more than 170 resources. Other network-based tools such as SLAP [6] and ProphNet [10] include 17 and 3 databases respectively. In terms of diversity, EKP includes databases that span drug, disease, phenotype, protein, gene and molecular pathways. Additionally, EKP takes advantage of mining the PubMed published literature. To our knowledge this is the most resource inclusive effort in network-based drug disease associations. Our methodology utilizes drug disease connections beyond the commonly used drug-targets-disease framework to expand the possibilities to include other semantic categories, such as drug-drug and disease-disease similarities, phenotypes, pathways, proteins and biological function annotations.

Conclusions

Computational efforts in drug discovery are gaining popularity for their ability to reduce the costs involved in drug development. Network-based approaches are currently being used for drug repurposing efforts. We have taken advantage of the EKP that integrates more than 170 biological sources. Leveraging 12 semantic categories that are found in the EKP to connect drug and disease pairs, we identified three main network features that showed significant differences in the characteristics of the intermediate concepts

connecting the drug disease pairs in the Gold Standard and negative datasets. These features can be readily used to build a classifier that will mine the full EKP graph to propose novel drug disease associations. Additional network features that are tailored to specific semantic types can be further extracted to fine tune the performance of the classifier.

This work demonstrates that semantic knowledge graphs have a strong potential in mitigating drug discovery efforts. We expect semantic graphs to grow with the exponential growth in data generation in life sciences. Thus, rendering semantic knowledge graphs even more valuable for drug discovery.

Table 1. Top 5 most significant semantic subcategories based on count feature.

Semantic type(subcategory)	Semantic Subcategory	Kolmogorov-Smirnov p-value
Organic Chemical	Chemicals & Drugs	7.39E-23
Pharmacologic Substance	Chemicals & Drugs	1.09E-22
Indicator, Reagent, or Diagnostic Ai	Chemicals & Drugs	7.12E-19
Hazardous or Poisonous Substance	Chemicals & Drugs	1.2E-16
Chemical Viewed Structurally	Chemicals & Drugs	2.72E-16

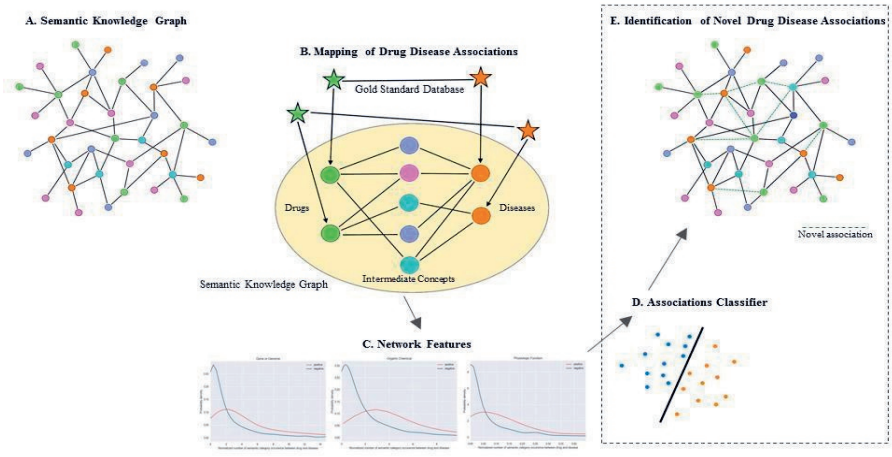

Fig. 1. We have used the Euretos Knowledge Platform (EKP) as the semantic knowledge graph in this analysis. Biological concepts (e.g. drugs, diseases, genes) are represented as circles, with different colors suggesting the variety of semantic types in the EKP (A). The drug disease pairs we acquired from an independent source were mapped to EKP concepts (B). Notably, mapped pairs were connected by intermediate concepts of 12 out of 14 different semantic types. We extracted network features from the intermediate concepts (C) to use them in building a classifier (future work) (D) to predict novel drug disease associations in the semantic network (E). Black dashed line reflects ongoing parts D and E of which their results are not yet included in this manuscript.

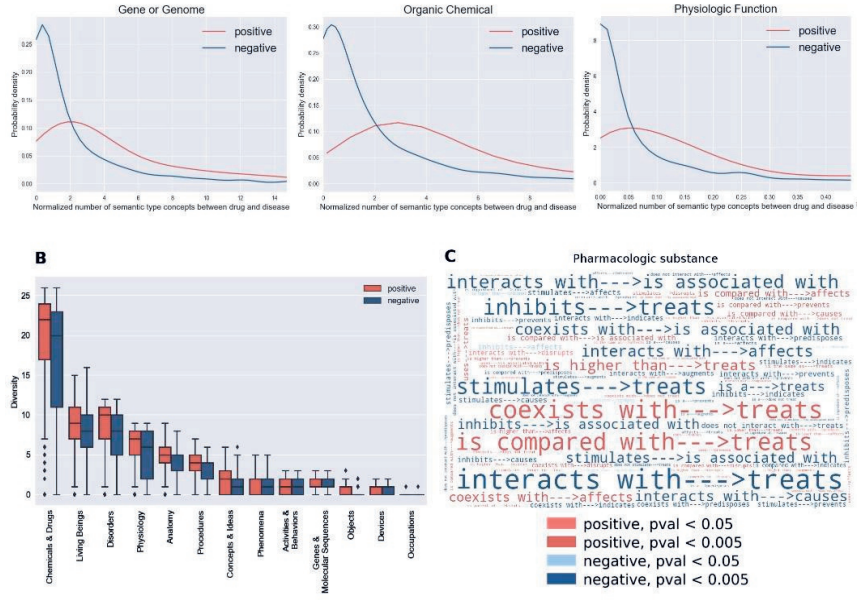

Fig. 2. A). Density plots of the count feature of three semantic subcategories. The higher the count value on the x-axis the higher the higher this semantic subcategory is found as an intermediate concept between drug and disease pairs. B). Boxplots representing the diversity feature for concepts in each of the 12 semantic categories. For each semantic category, we have calculated the presence of each of the subcategories belonging to that semantic category. C). Word Cloud representation of predicate paths of the "Pharmacological Substance" semantic category. P-values of the chi-square test residuals were used as an input to the cloud to calculate the enrichment of each path in either the positive and the negative datasets.

Acknowledgments

The research leading to these results has received funding from the People Program (Marie Curie Actions) of the European Union's Seventh Framework Program FP7/2077-2013 under REA grant agreement no. 317246. In addition, the European Commission (FP-7 project RD-Connect, grant agreement No. 305444).

Competing Interests

Kristina M. Hettne has performed paid consultancy since November 1, 2015, for Euretos b.v, a startup founded in 2012 that develops knowledge management and discovery services for the life sciences, with the Euretos Knowledge Platform as a marketed product

References

1. Choi, S., Macalino, S.J.Y., Cui, M., Basith, S.: Expediting the Design, Discovery, and Development of Anticancer Drugs using Computational Approaches. Curr. Med. Chem. (2016).
2. Glick, M., Jacoby, E.: The role of computational methods in the identification of bioactive compounds. Curr. Opin. Chem. Biol. 15, 540–546 (2011).

3. Perlman, L., Gottlieb, A., Atias, N., Ruppin, E., Sharan, R.: Combining drug and gene similarity measures for drug-target elucidation. J. Comput. Biol. 18, 133–145 (2011).
4. Yildirim, M.A., Goh, K.-I., Cusick, M.E., Barabási, A.-L., Vidal, M.: Drug-target network. Nat. Biotechnol. 25, 1119–1126 (2007).
5. Wu, Z., Wang, Y., Chen, L.: Network-based drug repositioning. Mol. Biosyst. 9, 1268–1281 (2013).
6. Chen, B., Ding, Y., Wild, D.J.: Assessing drug target association using semantic linked data. PLoS Comput. Biol. 8, e1002574 (2012).
7. Hettne, K.M., Thompson, M., van Haagen, H.H.H.B.M., van der Horst, E., Kaliyaperumal, R., Mina, E., Tatum, Z., Laros, J.F.J., van Mulligen, E.M., Schuemie, M., Aten, E., Li, T.S., Bruskiewich, R., Good, B.M., Su, A.I., Kors, J.A., den Dunnen, J., van Ommen, G.-J.B., Roos, M., 't Hoen, P.A.C., Mons, B., Schultes, E.A.: The Implicitome: A Resource for Rationalizing Gene-Disease Associations. PLoS One. 11, e0149621 (2016).
8. Guney, E., Menche, J., Vidal, M., Barábasi, A.-L.: Network-based in silico drug efficacy screening. Nat. Commun. 7, 10331 (2016).
9. McCray, A.T., Burgun, A., Bodenreider, O.: Aggregating UMLS semantic types for reducing conceptual complexity. Stud. Health Technol. Inform. 84, 216–220 (2001).
10. Martínez, V., Cano, C., Blanco, A.: ProphNet: a generic prioritization method through propagation of information. BMC Bioinformatics. 15 Suppl 1, S5 (2014).

CHAPTER 6
DISCUSSION

1. Introduction to molecular signatures

In recent years, new high-throughput measurement technologies for biomolecules such as DNA, RNA, and proteins have enabled unprecedented views of biological systems at the molecular level. Molecular signatures are based on *omics* data, made of a single or multiple data types (genome, transcriptome, proteome, and metabolome data) used to identify biological signals. Molecular signatures could be used to predict a phenotype of clinical interest such as a cancerous or diabetic state, molecules not previously known to be involved in the underlying molecular pathology, or molecular response to therapeutic drugs and their physiological toxicity. In this thesis, we used transcriptomic signatures to discover the molecular events that lead to the formation of cysts in autosomal dominant polycystic kidney disease and promote novel drug candidates for its treatment. There are countless examples of successful use of molecular signatures in disease study and drug discovery. For instance, the work of *Angus et al* demonstrates the application of whole-genome sequencing to identify mutational signatures linked to different types of chemotherapy pretreatments for breast cancer[1], thus providing useful genomic insights that could be leveraged to improve the future of patient management.

2. Robustness of molecular signatures and use in disease study

Despite their wide use, molecular signatures' success is hindered by limited reproducibility and variable performance on independent test sets[2], which in a large part could be attributed to the low signal-to-noise ratio inherent to omics datasets, the prevalence of batch effects in omics data, and molecular heterogeneity between samples and within populations[3]. The noise could be due to errors that are technical in nature (uncontrollable variation in different instrument readings collected from the same sample) or biological in nature (uncontrollable variation in different samples collected from the same biological condition). An event may also be considered part of noise even if it is biological and reproducible, simply because it encodes aspects of phenotype irrelevant to the current study. These issues are exacerbated by the fact that the datasets used to develop molecular signatures tend to have small sample sizes relative to the number of molecular measurements[4,5]. To overcome these limitations, experimentalists are deploying several strategies to enhance true biological signal detection. Meta-analysis of experimental data is an example of these methods. This approach combines the molecular signature of single studies into a uniform dataset that greatly eliminates single study anomalies and enables the identification of a refined molecular signature that limits the influence of experimental design, choice of the disease model, analysis methods, etc. on the biological signal. For instance, in **chapter-2** of this thesis, we combined four ADPKD expression profiling studies by focusing on genes that are significantly dysregulated in two or more studies, which led to the identification of the PKD Signature, a set of 1,515 genes that are commonly dysregulated in PKD studies. Our work revealed that genes that appear consistently dysregulated in multiple studies are significantly more enriched when tested against an independent relevant study, and are less prone to experimental differences. We also combined 7 renal injury-induced studies to eliminate noise and identify genes implicated in renal injury repair and involved in ADPKD pathology. Another method to eliminate noise in biological data relies on filtering biological events using prior knowledge. Prior knowledge could be stored in biological databases such as pathway annotation databases (e.g., KEGG), or stored in public literature (e.g., PubMed). In **chapter-**

2, we used several databases like KEGG, MSigDB and DAVID to validate the identified ADPKD signature, by measuring the enrichment of ADPKD associated pathway genes (e.g., proliferation, apoptosis, immune response, cell cycle) in the ADPKD signature and established that the ADPKD signature had twice the enrichment for ADPKD associated pathway genes compared to signatures derived from single studies. We applied a similar approach in **chapter-3**, where we used prior annotation of transcription factors (TFs) in MSigDB[6] to identify TFs associated with ADPKD (e.g., STAT3, and RUNX1) on the basis of their dysregulation in the ADPKD signature. Furthermore, we demonstrated the utility of leveraging public literature to filter noise and enhance true biological signals by mining scientific knowledge from PubMed abstracts[7] in the identification of injury-repair-associated genes in PKD in **chapter-2**. Using this method, we identified 237 injury repair-associated genes from the literature to supplement the injury profile we identified from expression data.

A third technique used by scientists to enhance signal-to-noise ratio relies on integrating complementary layers of data. For instance, scientists combine genomics and transcriptomics, to gain insights into the expression patterns of genes with somatic mutations. This method is extremely helpful in understanding the molecular events contributing to cancerous states. In addition, combining transcriptomics and proteomics data has been shown to improve disease study and drug development[8-10]. This is demonstrated by the efforts of Varemo et al that integrated transcriptomics and proteomics data to reconstruct the Human Myocyte Metabolic Network for the identification of Diabetes Markers[11]. Other forms of molecular data were also combined, for instance, transcriptomics and metabolomics[12], proteomics and metabolomics[13] and Chromatin immunoprecipitation with DNA sequencing (ChIP-Seq)[14]. In **chapter-3**, we identified STAT3 and RUNX1 as TFs with altered expression in the PKD Signature, and combining transcriptomics data with ChIP-Seq data we identified several key target protein for these TFs; STAT3 (*Scp2, Kif22 and Socs3*) and RUNX1 (*Bcl3, Tnfrsf12a*) which could be further investigated for their role in PKD development.

3. Molecular signatures for drug repurposing and discovery

In silico drug discovery methods can be classified into three types. The first is based on transcriptional molecular signatures, which measure genome-wide transcriptional perturbations of a biological system after drug induction, leading to the identification of the signature of the compounds' activity on biological systems. These molecular transcriptional signatures can then be compared to establish therapeutic relationships between known drugs and new disease indications. In **chapter-4** for instance, we have first identified the ADPKD disease progression profile by sequencing the transcriptome of *Pkd1cko* mice with varying disease severity (mild, moderate, and advanced). Furthermore, we identified significantly dysregulated genes in *Pkd1cko* mice treated with drugs proven to reduce cyst size and slow ADPKD progression in murine (Rapamycin (Sirolimus), Curcumin, and soluble activin receptor IIB Fc (sActRIIB-Fc)). Combining the ADPKD progression profile genes with the genes altered in the drug-treated mice, we identified genes that could potentially be targeted to modulate ADPKD progression. Although the approach we followed would not identify the molecular aspects of the drugs' cyst reducing mechanism, it allowed us to identify the transcriptome of the milder forms of ADPKD and consequently the genes that are implicated with ADPKD's progression. However, the exact role of these differentially expressed genes in ADPKD progression is difficult to predict. Further analysis of these

genes via the ChEMBL database led to the identification of 116 drug candidates targeting 29 proteins, of which several were previously linked to renal disease such as Rosiglitazone[15]. One of the most comprehensive and systematic approaches leveraging this approach is the LINCS database, previously known as the Connectivity Map project[16]. Initially, the LINCS database contained profiles of 164 drugs and currently has a library containing over 1.5M gene expression profiles from ~5,000 small-molecule compounds, and ~3,000 genetic reagents, tested in multiple cell types. These signatures then form the basis of comparing drugs' mechanism of action at the transcriptional level.

The second class of methods is network-based that aims at organizing the relationships among biological molecules in the form of networks to find newly emerged properties at a network level, and to investigate how cellular systems induce different biological phenotypes under different conditions. A network can be depicted as a connected graph, where each node can represent either an individual molecular entity (e.g., a drug), its biological target, a modifier molecule within a biological process, or a target pathway, while an edge represents either a direct or indirect interaction between two connected nodes. By applying novel algorithms on these networks, novel associations between drugs and diseases, and drugs and protein targets could be derived allowing for the discovery of novel drug candidates to treat various diseases. In **chapter-5,** we analyzed semantic knowledge graphs, a subclass of knowledge networks that include data mined from literature, to demonstrate the utility of network-based approaches in identifying drug-disease with potential clinical efficacy. Our results revealed that the extracted network-features such as the total number of intermediate concepts (count), the number of different semantic categories (diversity), and the predicates connecting a drug-disease pair were successful in separating the positive from the negative sets. The positive dataset included 239 drugs, 78 diseases, and 403 drug-disease pairs[17] and their drug-disease interactions were analyzed within the Euretos Knowledge Platform, a commercial knowledge graph that semantically integrates 200 biomedical knowledge sources. This work demonstrates that semantic knowledge graphs have a strong potential in mitigating drug discovery efforts.

The third class of methods is ligand-based and assumes similar compounds tend to have similar biological properties, and similarity between representations is used to assess whether a compound modulates the activity of a target or treats a disease like a known drug[18]. Such methods make use of large-scale virtual screening experiments that analyze molecular shape or molecular docking data to suggest possible further development of hits into lead[19,20]. These methods are becoming more feasible given the drastic growth in public databases, such as PubChem, ChEMBL, and DrugBank, which store a large amount of chemical/biological information such as cellular activity, binding activity, and functional data[21-23]. For instance, in **chapter-4** we screened the ChEMBL database for all clinically tested drugs (only drugs tested in phases 2, 3, or 4 of clinical trials) that target the proteins we identified as part of the ADPKD progression profile. As we anticipate that ADPKD treatment will be administered in the long-term and to limit drugs that might cause serious side effects, we filtered all drugs with antineoplastic classification based on the Anatomical Therapeutic Chemical (ATC) Classification System. In addition to the basic aforementioned methods, a new set of methods that utilize machine learning are gaining popularity, especially relevant for identifying disease-drug associations. For instance, Zhang et al introduced a matrix factorization method that combines bioactivity data with disease

information, in one framework to identify new drug candidates based on drug similarity and disease similarity[24].

Additionally, computational tools are increasingly being leveraged to predict drug-target interactions to narrow down the scope of the search of candidate medications [25]. The majority of these methods rely on the knowledge of the three-dimensional structure of the biomolecular target to estimate the affinity of ligands to a target protein with considerable accuracy and efficiency, however, the availability of 3D protein structures remains a significant bottleneck[26].

4. Future outlook, exciting developments, and key challenges in drug discovery

Drug discovery and development is a time-consuming and extremely expensive process. It is estimated that the research and development cost for one molecular entity is approximately USD 2.7 billion[27]. In recent years, the drug discovery process has benefited from the latest technological developments in biomedicine. In particular, the increased availability of high-throughput omics technologies resulted in a wealth of biological data that we are only starting to tap into. Data availability has drastically increased with the decline in sequencing costs, enabling higher data resolution (e.g., single cells or subcellular localizations). For instance, Single-cell transcriptome profiling (scRNA-seq) has enabled high-resolution mapping of cellular heterogeneity, development, and activation states in diverse systems, furthering our understanding of disease pathogenesis[28,29]. We are also witnessing an increase in the type of molecular data available to study diseases enabling a systems-view into disease etiology across DNA, RNAs, and proteins. For instance, *cis*-regulatory elements and non-coding RNAs are shown to play an important role in disease development[30,31]. The phenome is another source of data that is exploited recently for drug discovery, and phenome-wide association studies (PheWAS) are gaining popularity as a systematic approach to statistically estimate the association between single-nucleotide polymorphisms and a large number of different phenotypes[32]. The work of *Denny et al* demonstrated the viability of PheWAS (based on EMRs) in detecting novel associations between genetic markers and human diseases[33].

In addition to the growth in data diversity and availability, there are major developments in experimental technologies to study diseases and accelerate drug discovery. A key example is the use of CRISPR/Cas9-based methods to perform combinatorial screens effectively in cells and develop therapeutics. In drug discovery, CRISPR/Cas9 allows us to perform genome-wide screening by systematically inactivating, or knocking out, the ~20,000 protein-coding genes found in humans to discover which proteins should be targeted for a disease treatment[34,35]. CRISPR/Cas9 could also be used in therapeutic applications by allowing scientists to target the underlying cause of disease and possibly cure it by modifying the patient's genome[36,37]. Organoids (3-D culture systems) are another technology expected to develop into a powerful tool for drug discovery. Organoids are based on ex vivo

biopsy samples from patients or animal models used for drug testing, gene editing, or research on prognosis. Organoids could be used to mimic animal or human tissue in a dish, from its early development to its organogenesis or adult stage. For instance, growing human-derived tissues ex vivo has opened the possibility to study human development, to model human disease directly from individual patients, or to test therapeutic compounds in a personalized medicine approach [38]. In fact, in **chapter-4** we used cyst-like organoids derived from *Pkd1*-KO mouse-inner medullary collecting duct (mIMCD3) cells to test the impact of computationally prioritized drug candidates on cysts formation.

Although these technologies provide immense hope for biomedical research, several challenges need to be tackled. The increase in data resolution and data types increases the complexity of integrating them and requires the use of automated methods to process large datasets rapidly and robustly, and then accurate models to identify meaningful insights such as drivers of gene expression patterns in time and across organs upon perturbation. Furthermore, data security, sharing, and ownership issues will become essential factors of biomedical research.

REFERENCES

1 Angus, L. *et al.* The genomic landscape of metastatic breast cancer highlights changes in mutation and signature frequencies. *Nat Genet* **51**, 1450-1458, doi:10.1038/s41588-019-0507-7 (2019).

2 Sung, J., Wang, Y., Chandrasekaran, S., Witten, D. M. & Price, N. D. Molecular signatures from omics data: from chaos to consensus. *Biotechnol J* **7**, 946-957, doi:10.1002/biot.201100305 (2012).

3 Ideker, T., Dutkowski, J. & Hood, L. Boosting signal-to-noise in complex biology: prior knowledge is power. *Cell* **144**, 860-863, doi:10.1016/j.cell.2011.03.007 (2011).

4 Dougherty, E. R. Small sample issues for microarray-based classification. *Comp Funct Genomics* **2**, 28-34, doi:10.1002/cfg.62 (2001).

5 Stretch, C. *et al.* Effects of sample size on differential gene expression, rank order and prediction accuracy of a gene signature. *PLoS One* **8**, e65380, doi:10.1371/journal.pone.0065380 (2013).

6 Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst* **1**, 417-425, doi:10.1016/j.cels.2015.12.004 (2015).

7 Jelier, R. *et al.* Anni 2.0: a multipurpose text-mining tool for the life sciences. *Genome Biol* **9**, R96, doi:10.1186/gb-2008-9-6-r96 (2008).

8 Eddy, J. A., Hood, L., Price, N. D. & Geman, D. Identifying tightly regulated and variably expressed networks by Differential Rank Conservation (DIRAC). *PLoS Comput Biol* **6**, e1000792, doi:10.1371/journal.pcbi.1000792 (2010).

9 Li, C. & Li, H. Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics* **24**, 1175-1182, doi:10.1093/bioinformatics/btn081 (2008).

10 Witham, F. H. & Hendry, L. B. Computer modeling of gibberellin-DNA binding. *J Theor Biol* **155**, 55-67, doi:10.1016/s0022-5193(05)80548-x (1992).

11 Varemo, L. *et al.* Proteome- and transcriptome-driven reconstruction of the human myocyte metabolic network and its use for identification of markers for diabetes. *Cell Rep* **11**, 921-933, doi:10.1016/j.celrep.2015.04.010 (2015).

12 Hirai, M. Y. *et al.* Integration of transcriptomics and metabolomics for understanding of global responses to nutritional stresses in Arabidopsis thaliana. *Proc Natl Acad Sci U S A* **101**, 10205-10210, doi:10.1073/pnas.0403218101 (2004).

13 Yizhak, K., Benyamini, T., Liebermeister, W., Ruppin, E. & Shlomi, T. Integrating quantitative proteomics and metabolomics with a genome-scale metabolic network model. *Bioinformatics* **26**, i255-260, doi:10.1093/bioinformatics/btq183 (2010).

14 Hull, R. P. *et al.* Combined ChIP-Seq and transcriptome analysis identifies AP-1/JunD as a primary regulator of oxidative stress and IL-1beta synthesis in macrophages. *BMC Genomics* **14**, 92, doi:10.1186/1471-2164-14-92 (2013).

15 Liu, C., Zhang, Y., Yuan, L., Fu, L. & Mei, C. Rosiglitazone inhibits insulin-like growth factor1-induced polycystic kidney disease cell growth and p70S6 kinase activation. *Mol Med Rep* **8**, 861-864, doi:10.3892/mmr.2013.1588 (2013).

16 Lamb, J. *et al.* The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* **313**, 1929-1935, doi:10.1126/science.1132939 (2006).

17 Guney, E., Menche, J., Vidal, M. & Barabasi, A. L. Network-based in silico drug efficacy screening. *Nat Commun* **7**, 10331, doi:10.1038/ncomms10331 (2016).

18 Schuler, J. & Samudrala, R. Fingerprinting CANDO: Increased Accuracy with Structure- and Ligand-Based Shotgun Drug Repurposing. *ACS Omega* **4**, 17393-17403, doi:10.1021/acsomega.9b02160 (2019).

19 da Silva Rocha, S. F. L., Olanda, C. G., Fokoue, H. H. & Sant'Anna, C. M. R. Virtual Screening Techniques in Drug Discovery: Review and Recent Applications. *Curr Top Med Chem* **19**, 1751-1767, doi:10.2174/1568026619666190816101948 (2019).

20 Ma, X. H. *et al.* Virtual screening methods as tools for drug lead discovery from large chemical libraries. *Curr Med Chem* **19**, 5562-5571, doi:10.2174/092986712803833245 (2012).

21 Kim, S. *et al.* PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res* **47**, D1102-D1109, doi:10.1093/nar/gky1033 (2019).

22 Gaulton, A. *et al.* ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res* **40**, D1100-1107, doi:10.1093/nar/gkr777 (2012).

23 Wishart, D. S. *et al.* DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res* **36**, D901-906, doi:10.1093/nar/gkm958 (2008).

24 Zhang, P., Wang, F. & Hu, J. Towards drug repositioning: a unified computational framework for integrating multiple aspects of drug similarity and disease similarity. *AMIA Annu Symp Proc* **2014**, 1258-1267 (2014).

25 Anusuya, S. *et al.* Drug-Target Interactions: Prediction Methods and Applications. *Curr Protein Pept Sci* **19**, 537-561, doi:10.2174/1389203718666161108091609 (2018).

26 Wang, X., Song, K., Li, L. & Chen, L. Structure-Based Drug Design Strategies and Challenges. *Curr Top Med Chem* **18**, 998-1006, doi:10.2174/1568026618666180813152921 (2018).

27 DiMasi, J. A., Grabowski, H. G. & Hansen, R. W. Innovation in the pharmaceutical industry: New estimates of R&D costs. *J Health Econ* **47**, 20-33, doi:10.1016/j.jhealeco.2016.01.012 (2016).

28 Szabo, P. A. *et al.* Single-cell transcriptomics of human T cells reveals tissue and activation signatures in health and disease. *Nat Commun* **10**, 4706, doi:10.1038/s41467-019-12464-3 (2019).

29 Park, J. *et al.* Single-cell transcriptomics of the mouse kidney reveals potential cellular targets of kidney disease. *Science* **360**, 758-763, doi:10.1126/science.aar2131 (2018).

30 Lee, H., Zhang, Z. & Krause, H. M. Long Noncoding RNAs and Repetitive Elements: Junk or Intimate Evolutionary Partners? *Trends Genet* **35**, 892-902, doi:10.1016/j.tig.2019.09.006 (2019).

31 Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **22**, 1760-1774, doi:10.1101/gr.135350.111 (2012).

32 Hebbring, S. J. The challenges, advantages and future of phenome-wide association studies. *Immunology* **141**, 157-165, doi:10.1111/imm.12195 (2014).

33 Denny, J. C. *et al.* Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat Biotechnol* **31**, 1102-1110, doi:10.1038/nbt.2749 (2013).

34 Shi, J. *et al.* Discovery of cancer drug targets by CRISPR-Cas9 screening of protein domains. *Nat Biotechnol* **33**, 661-667, doi:10.1038/nbt.3235 (2015).

35 Shalem, O. *et al.* Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science* **343**, 84-87, doi:10.1126/science.1247005 (2014).

36 Lim, K. R. Q., Yoon, C. & Yokota, T. Applications of CRISPR/Cas9 for the Treatment of Duchenne Muscular Dystrophy. *J Pers Med* **8**, doi:10.3390/jpm8040038 (2018).

37 Pankowicz, F. P. *et al.* Reprogramming metabolic pathways in vivo with CRISPR/Cas9 genome editing to treat hereditary tyrosinaemia. *Nat Commun* **7**, 12642, doi:10.1038/ncomms12642 (2016).

38 Huch, M., Knoblich, J. A., Lutolf, M. P. & Martinez-Arias, A. The hope and the hype of organoid research. *Development* **144**, 938-941, doi:10.1242/dev.150201 (2017).

APPENDIX

Nederlandse Samenvatting

In de afgelopen 20 jaar zijn de medische wetenschappen getuige geweest van belangrijke resultaten die werden gekenmerkt door de voltooiing van het menselijk genoomproject in 2003. Vanaf dat moment daalden de kosten van DNA-sequentiebepaling aanzienlijk en dit markeerde vervolgens de komst van het data-tijdperk in de medische wetenschappen. In dit proefschrift hebben we me gericht op het onderzoeken van de voordelen van big data bij de behandeling van ziekten bij de mens, en specifiek op het onderzoeken van nieuwe behandelingen voor autosomaal dominante polycysteuze nierziekte (ADPKD). ADPKD is een ziekte die de nieren van de patiënt aantast en kenmerkt zich door met vloeistof gevulde cysten die zich in beide nieren vormen. ADPKD heeft momenteel zeer beperkte behandelingsopties; patiënten kunnen nierfalen ontwikkelen en hebben uiteindelijk vaak een niertransplantatie nodig. In hoofdstuk 2 van dit proefschrift zijn we begonnen met het combineren van de verschillende openbaar beschikbare expressieprofielen van ADPKD om te komen tot een uniform ADPKD-genexpressie profiel. Expressieprofielen worden gemaakt door de volledige set RNA-transcripten (tot expressie gebrachte genen) onder specifieke omstandigheden of in een specifieke cel vast te leggen - met behulp van high-throughput sequencing-methoden. Het hebben van een uniforme expressieprofiel vergroot ons begrip van de genen die bij de ziekte betrokken zijn en minimaliseert fouten die zouden kunnen voortkomen uit het gebruik van verschillende technologieën of verschillende diermodellen. Enkele experimenten zijn beperkt tot een specifieke technologie, diermodel en / of ziektetoestand. Het is waardevol om verschillende onderzoeken te combineren om deze beperkingen te overwinnen. In ons werk hebben we aangetoond dat het combineren van experimenten met verschillende ziektemodellen en sequentie technologieën een sterker expressieprofiel creëert met een grotere percentage van genen die betrokken zijn bij de ziekte. Verder hebben we in hoofdstuk 3 deze expressie-signatuur gebruikt om een klasse van genen genaamd transcriptiefactoren (TF's) en hun rol in ADPKD te bestuderen. Transcriptiefactoren zijn bijzonder interessant omdat ze in staat zijn de expressieniveaus van andere genen te reguleren. Het zijn daarmee belangrijke organisatoren van moleculaire gebeurtenissen die aanleiding kunnen geven tot de ontwikkeling van cysten bij ADPKD-patiënten. In hoofdstuk 3 ontdekten we twee TF's, STAT3 en RUNX1, die verder onderzocht zouden kunnen worden op hun rol in cystenvorming. We hebben ook een paar andere genen belicht die door deze twee TF's worden gereguleerd en als een mogelijk aangrijpingspunt kunnen dienen voor medicijnen voor ADPKD-patiënten. In hoofdstukken 2 en 3 combineerden we grote hoeveelheden transcriptomics-gegevens, mogelijk gemaakt door aanzienlijke vooruitgang in DNA-sequentie technologieën.

Verder hebben we in hoofdstuk 4 de sequentie-analyses uitgevoerd voor 24 nierweefsels uit ons ADPKD muismodel (Pkd1cko) in verschillende ziektestadia om de genen te bestuderen die betrokken zijn bij cystevorming en -progressie. We hebben de nieren op basis van de ernst van de ziekte in drie groepen ingedeeld; Vroege fase met nog bijna geen vorming van cysten, intermediaire fase met zichtbare cysten en eindfase met grote cysten die zich in de nieren vormen. We combineerden de gegevens van deze drie ziektefasen om te onderzoeken hoe veranderingen in genexpressie verband houden met de progressie van ADPKD, en kwam tot een reeks genen die betrokken zouden kunnen zijn bij de progressie van ADPKD, genaamd "ADPKD progressie profiel". Verder hebben we de chemo-informatica-database (ChEMBL) gebruikt om te identificeren welke medicijnen kunnen

worden gebruikt om belangrijke genen in het “ADPKD progressie profiel” te targeten en zo de vorming van cysten en ziekteprogressie te vertragen of te stoppen. Dit werk nomineerde verschillende nieuwe kandidaat-geneesmiddelen om ADPKD te behandelen, zoals Birinapant en Gamma-linoleenzuur. In dit hoofdstuk hebben we specifiek gekeken naar geneesmiddelen die al in de klinische praktijk zijn goedgekeurd om andere ziekten te behandelen. Het onderzoeken van het gebruik van een medicijn dat voor een bepaalde ziekte wordt gebruikt voor een andere indicatie, is een opkomend gebied van medicijnontdekking, genaamd drug repurposing. Drug repurposing zou het ontwikkel- en toelatingstraject van medicijnen aanzienlijk kunnen versnellen. Er wordt gebruik gemaakt van onze klinische en toxicologische kennis van huidige medicijnen, waardoor ze gemakkelijker toe te passen zijn voor andere indicaties dan onbekende medicijnen. Daarnaast hebben we in dit hoofdstuk bio-informatica gecombineerd die zich bezighoudt met gegevens van genen en eiwitten, met chemo-informatica die zich bezighoudt met gegevens met betrekking tot kleine moleculen en medicijnen.

Tenslotte richtte ik me in hoofdstuk 5 op de behoefte aan methoden die de enorme hoeveelheden biologische gegevens die dagelijks worden geproduceerd, kunnen combineren tot een zinvol formaat voor machines om wetenschappers te helpen nieuwe waardevolle relaties in de biologische wetenschappen te vinden. Kennis netwerken zijn een steeds populairder formaat om biologische gegevens op te slaan, waarbij gegevens uit verschillende biologische bronnen worden gecombineerd en geannoteerd tot betekenisvolle relaties die snel door machines kunnen worden doorzocht. Door al deze databronnen op één plek te hebben, kunnen wetenschappers verbanden leggen tussen verschillende vormen van data en komen veel biologische vragen aan de orde, zoals "welke medicijnen kunnen worden gebruikt om kanker te behandelen?", Of "welke genen veroorzaken diabetes?". In hoofdstuk-5 van dit proefschrift hebben we onderzocht hoe we dergelijke kennis netwerken kunnen explorerenmet behulp van machine learning om te voorspellen welke biologische relaties “waar” en van hoge kwaliteit zouden kunnen zijn, zodat ze verder kunnen worden getest door wetenschappers in het laboratorium. Dergelijke methoden zijn erg belangrijk om biologische ontdekkingen te versnellen en fouten te minimaliseren. Het belang hiervan zal toenemen naarmate we doorgaan met het produceren van enorme hoeveelheden biologische gegevens die ons vermogen om te interpreteren en om te zetten in zinvolle biologische toepassingen te boven gaan.

English Summary

In the past 20 years the field of medical sciences witnessed significant achievements marked by the completion of the human genome project in 2003. From that moment, the costs of DNA sequencing dropped significantly. This marked the arrival of the data era in medical sciences. In this thesis, we focused on exploring the benefits of big data in treating human diseases, and specifically investigating novel treatments for Autosomal Dominant Polycystic Kidney Disease (ADPKD). ADPKD is a disease that impacts the patient's kidneys, marked by liquid-filled cysts forming in both kidneys. ADPKD currently has very limited treatment options and could lead to patients developing end-stage renal disease and ultimately requiring kidney transplant.

In chapter-2 of this thesis, we started by combining the different publicly available expression profiles of ADPKD, to arrive at a uniform ADPKD gene expression signature. Expression profiles are created by capturing the complete set of RNA transcripts (expressed genes) under specific circumstances or in a specific cell, using high-throughput sequencing analysis methods. Having a uniform expression signature enhances our understanding of the genes that are involved in the disease and minimizes errors that could come from using different technologies or different animal models when studying the disease. Single experiments are limited to a specific technology, animal model and/or disease state, there is value in combining different studies to overcome these limitations. In our work, we demonstrated that combining experiments across disease models and sequencing technologies creates a stronger gene signature with a higher chance of including genes that are modulated by the disease. Furthermore, in chapter-3, we used this expression signature to study a class of genes called transcription factors (TFs) and their role in ADPKD. Transcription factors are particularly interesting as they are capable of regulating the expression levels of other genes, making them key orchestrators of molecular events that could give rise to events such as cysts development in ADPKD patients. In chapter-3, we discovered two TFs, STAT3 and RUNX1 that could be further explored for their role in cysts formation. We also highlighted a few other genes that could be drug targeted by these two TFs to form a network of genes that act together to lead to significant molecular changes in ADPKD patients. In both chapters- 2 and 3, we relied on our ability to combine large amounts of transcriptomics data made possible by significant advancements in DNA sequencing technologies.

Furthermore, in chapter-4, we sequenced 24 kidney tissues taken from our ADPKD mouse model (Pkd1cko) at different disease stages to study the genes implicated in cyst formation and progression. We grouped the kidneys based on their disease severity into three groups; Early-phase with almost no cysts formation yet, moderate-phase with visible cysts and, advanced-phase with large cysts formed in the kidneys. We combined the data from these three disease phases to explore how gene expression changes relate with the progression of ADPKD and were able to arrive at a set of genes that could be implicated in ADPKD progression called "ADPKD Progression Profile". Furthermore, we leveraged the chemoinformatics database (ChEMBL) to identify which drugs could be used to target important genes in the "ADPKD Progression Profile", and hence slow down or stop cyst formation and disease progression. Combining bioinformatics that is concerned with data from genes and protein with chemoinformatics that is concerned with data related to small molecules and drugs, enables a detailed analysis of which drugs could be used to target important proteins involved with ADPKD progress.This work nominated several novel drug candidates to treat ADPKD, such as Birinapant and Gamma-Linolenic acid. In this chapter, we looked specifically at drugs that are already approved in clinical practice to treat other

diseases. Exploring the use of a drug used for a certain disease for another indication is an emerging field of drug discovery called drug repurposing. Drug repurposing could significantly speed up drug development and market approval because it can leverage on existing clinical and toxicological knowledge.

Finally, in chapter-5, we focused on addressing the need to have tools that can combine the vast amounts of biological data being produced on a daily basis into a meaningful format for machines to help scientists find new valuable relationships in biology. Knowledge graphs is an increasingly popular format to store biological data, where data combined from different biological sources, and annotated into meaningful relationships that could be quickly searched by machines. Having all these data sources in a single place, enables scientists to draw links between different forms of data and to address many biological questions such as “which drugs can be used to treat cancer?”, or “which genes cause diabetes?”. In chapter-5 of this thesis, we studied how we could mine such knowledge graphs using machine learning to predict which biological relationships could be “true” and of high quality so that they could be further tested by scientists in the lab. Such methods are very important to speed-up biological discoveries and minimize errors. The importance of this kind of method will grow even further as we continue to produce vast amounts of biological data that outpaces our ability to interpret and convert into useful biological applications.

In chapter-6, we offer a summary of the key findings of this thesis and discuss how they fit in the overall scientific picture of advancing molecular biology and disease study. We highlight that in addition to transcriptomics, scientists are increasingly exploring additional sources of data such as genomics and proteomics to study the flow of genetic information, and metabolomics and lipidomics to study the products of metabolism. Additionally, we shed light on several challenges that need to be tackled to unlock the full potential of these technologies in biomedical research, such as the need for automated methods to quickly connect data across different data formats and help scientists derive useful biological connections. To conclude, we are living in an exciting era, where biological advancements have the potential to revolutionize our way of living, not just in healthcare, but also in other areas such as the food we eat, and the products we consume.

Curriculum Vitae

Tareq Malas was born on the 20th of August 1987 in Damascus, Syria. Tareq obtained his bachelor's degree in Information Systems at King Fahd University in Saudi Arabia with honors distinction in 2010. During his bachelor's studies, Tareq enrolled in few biological courses and enjoyed the study of cellular systems and decided to further explore the field of biosciences. In 2011, Tareq took several summer biosciences classes at the University of California San Diego and later that year won a full scholarship at King Abdullah University to complete his Master's degree in Biosciences. During his master's program, Tareq focused on combining his expertise in information systems with biosciences in the study of pathogen genomics. Under the supervision of Prof. Dr. Arnab Pain at King Abdullah University, Tareq was involved in several groundbreaking publications exploring the causative agents of several pathogenic diseases such as tuberculosis and coccidiosis. Tareq completed his Master's degree with distinction in 2012. In 2013, Tareq started his Ph.D. training by joining the lab of Prof. Dr. Dorien J.M. Peters at the Department of Human Genetics and the Biosemantics group led by Prof. Dr. Peter-Bram 't Hoen and Dr. Marco Roos at the Leiden University Medical Center to study the molecular aspects of Polycystic Kidney Diseases and explore feasible drug interventions to modulate disease progression. The findings obtained during his Ph.D. studies are described in this thesis. After his Ph.D. studies, Tareq joined McKinsey & Company in 2018 as a consultant working with major Life Sciences organizations on tackling key strategic topics such as the role of digital tools in advancing drug discovery and commercialization.

List of publications

Drug Repurposing Using a Semantic Knowledge Graph
Tareq B. Malas , Roman Kudrin, Sergei Starikov, Peter A.C.´t Hoen, Dorien J.M. Peters, Marco Roos, Kristina M. Hettne
SWAT4LS, Tech. Rep. 2017

Meta-analysis of polycystic kidney disease expression profiles defines strong involvement of injury repair processes
Tareq B Malas*, Chiara Formica*, Wouter N Leonhard, Pooja Rao, Zoraide Granchi, Marco Roos, Dorien J M Peters, Peter A C 't Hoen
Am J Physiol Renal Physiol. 2017;312(4):F806-F817. doi:10.1152/ajprenal.00653.2016

Early career researchers want Open Science
Andrea Farnham, Christoph Kurz, Mehmet Ali Öztürk, Monica Solbiati, Oona Myllyntaus, Jordy Meekes, Tra My Pham, Clara Paz, Magda Langiewicz, Sophie Andrews, Liisa Kanninen, Chantal Agbemabiese, Arzu Tugce Guler, Jeffrey Durieux, Sarah Jasim, Olivia Viessmann, Stefano Frattini, Danagul Yembergenova, Carla Marin Benito, Marion Porte, Anaïs Grangeray-Vilmint, Rafael Prieto Curiel, Carin Rehncrona, Tareq Malas, Flavia Esposito & Kristina Hettne
Genome Biol 18, 221 (2017). doi:10.1186/s13059-017-1351-7

Comprehensive transcriptome analysis of fluid shear stress altered gene expression in renal epithelial cells
Steven J. Kunnen, Tareq B. Malas, Cornelis M. Semeins, Astrid D. Bakker, Dorien J. M. Peters
J Cell Physiol. 2018; 233: 3615– 3628. doi:10.1002/jcp.26222

Nanopublications: A growing resource of provenance-centric scientific linked data
Tobias Kuhn, Albert Meroño-Peñuela, Alexander Malic, Jorrit H Poelen, Allen H Hurlbert, Emilio Centeno Ortiz, Laura I Furlong, Núria Queralt-Rosinach, Christine Chichester, Juan M Banda, Egon Willighagen, Friederike Ehrhart, Chris Evelo, Tareq B Malas, Michel Dumontier
2018 IEEE 14th International Conference on e-Science (e-Science), Amsterdam, 2018, pp. 83-92, doi: 10.1109/eScience.2018.00024. doi: 10.1109/eScience.2018.00024.

Comparative transcriptomics of shear stress treated Pkd1−/− cells and pre-cystic kidneys reveals pathways involved in early polycystic kidney disease
Steven J. Kunnen, Tareq B. Malas, Chiara Formica, Wouter N. Leonhard, Peter A.C.'t Hoen, Dorien J.M. Peters
Biomedicine & Pharmacotherapy, Volume 108, 2018, Pages 1123-1134, ISSN 0753-3322. doi: 10.1016/j.biopha.2018.07.178

Characterisation of transcription factor profiles in polycystic kidney disease (PKD): identification and validation of STAT3 and RUNX1 in the injury/repair response and PKD progression
Chiara Formica*, Tareq Malas*, Judit Balog, Lotte Verburg, Peter A. C. 't Hoen & Dorien J. M. Peters
J Mol Med 97, 1643–1656 (2019). doi: 10.1007/s00109-019-01852-3

* Contributed equally

Drug prioritization using the semantic properties of a knowledge graph
Tareq B. Malas*, Wytze J. Vlietstra*, Roman Kudrin, Sergey Starikov, Mohammed Charrout, Marco Roos, Dorien J. M. Peters, Jan A. Kors, Rein Vos, Peter A. C. 't Hoen, Erik M. van Mulligen & Kristina M. Hettne
Sci Rep 9, 6281 (2019). doi: 10.1038/s41598-019-42806-6

Prioritization of novel ADPKD drug candidates from disease-stage specific gene expression profiles
Tareq B. Malas, Wouter N. Leonhard, Hester Bange, Zoraide Granchi, Kristina M.Hettne, Gerard J.P. Van Westen, Leo S. Price, Peter A.C.'t Hoen, Dorien J.M. Peters
EBioMedicine, Volume 51, 2020, 102585, ISSN 2352-3964.
doi: 10.1016/j.ebiom.2019.11.046.

* Contributed equally

Acknowledgments

Foremost I would like to express my deepest gratitude to my Ph.D. supervisors, Prof. Dorien J.M. Peters and Prof. Peter-Bram 't Hoen. This thesis would have never been possible without your continuous support, guidance, and encouragement. Thank you for your patience with me and for reminding me to always give my utmost best to succeed. You welcomed me to be part of your research groups with a warm heart and showed me the way forward. I highly appreciate putting your trust and faith in me.

I also would like to thank the members of my PhD committee, Prof. Ko Willems van Dijk, Prof. Ingred Meulenbelt, Prof. Chris Evelo and Dr. Frank van Eeden.
A special thanks to my Ph.D. mentors Dr. Marco Roos, Dr. Kristina Hettne, and Dr. Wouter Leonhard; Your scientific expertise and wisdom contributed enormously to my success.

Furthermore, I would like to thank all of my colleagues at the PKD and Biosemantics groups. Our daily discussions, exchange of knowledge, mini celebrations when our papers got published, and our continued commitment to supporting each other, were a key motivation for me to cross the line.
A special mention to Prof. Gerard van Westen for helping me understand the value of chemoinformatics in drug research.

I am extremely grateful to my fellow TranCYST researchers, Chiara, Arianna, Martin, Kanishka, Alkaly, Laura, and Aylin. Our network events and social activities served as a major energizer to pull through the long nights of doing a Ph.D.

Last but not least, I would like to thank my parents, my sisters, my broader family and my friends. Mom and Dad, I can’t express how thankful I am to have you in my life, through your unconditional love I am able to overcome the most difficult of challenges.