Data-driven knowledge discovery in polycystic kidney disease

Malas, T.

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].

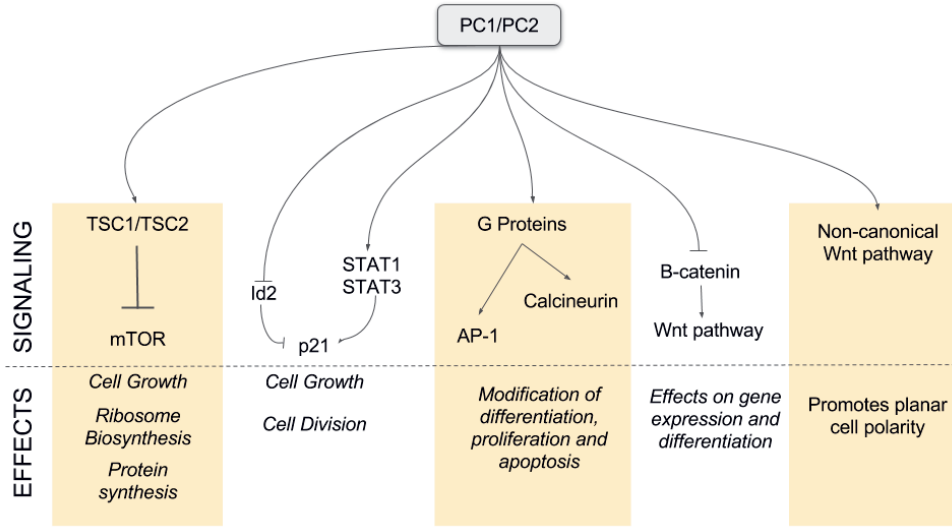

Figure-1: Overview of the signaling pathways affected by PC1 and PC2. PC1 and PC2 affect multiple signaling pathways. Figure-1 presents a summary of the signaling pathways influenced by PC1 and PC2. These pathways directly or indirect stimulate or inhibit different aspects of cellular growth and differentiation[51].

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.

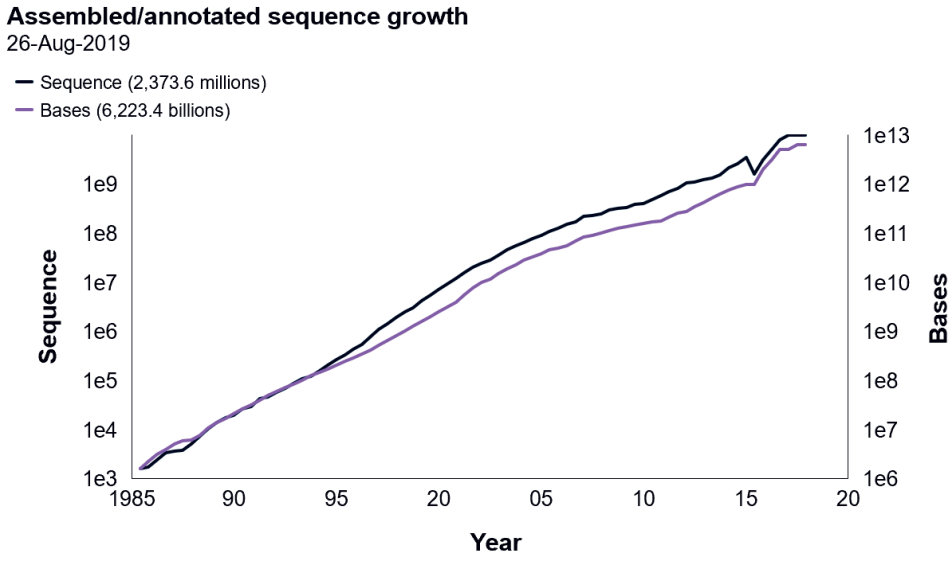

Figure-2: Number of sequences and bases deposited in the European Nucleotide Archive. The amount of sequences and DNA bases deposited in the European Nucleotide Archive – adopted from The European Nucleotide Archive in 2019.

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).