Department of Medical Biophysics and Biochemistry
Karolinska Institutet, Stockholm, Sweden

KIDNEY DISEASES: INSIGHTS FROM OMICS APPROACHES

Jing Guo

Karolinska Institutet

Stockholm 2018

Published by Karolinska Institutet.
Printed by Eprint AB 2018.

ISBN 978-91-7831-287-0

Kidney diseases: Insights from omics approaches

THESIS FOR DOCTORAL DEGREE (Ph.D.)

By

Jing Guo

Principal Supervisor:
Jaakko Patrakka
Karolinska Institutet
KI/AZ Integrated Cardio Metabolic Center
Department of Laboratory Medicine
Division of Pathology

Co-supervisor(s):
Karl Tryggvason
Karolinska Institutet
Department of Medical Biophysics and Biochemistry
Division of Matrix

Liqun He
Uppsala University
Department of Immunology, Genetics and Pathology
Division of Vascular Biology

Opponent:
Professor Matthias Kretzler
University of Michigan
Department of Internal Medicine
Division of Nephrology

Examination Board:
Professor Peter Hansell
Uppsala University
Department of Medical Cell Biology

Docent Sanna Lehtonen
University of Helsinki
Department of Pathology

Doctor Hong Jiao
Karolinska Institutet
Department of Biosciences and Nutrition

To my family

致我的家人

SUMMARY OF THE THESIS

Chronic kidney diseases (CKDs) affects about 11-15% of adults worldwide. When it progresses to the end-stage renal disease (ESRD), there is no effective medication for cure, the only treatment being chronic dialysis or kidney transplantation. The 5-year survival rate for patients in dialysis is less than 40%, and generates a huge economic burden to the healthcare system. A major proble is that we still have very limited knowledge on the pathogenesis and pathomechanism of CKD.

In this thesis, we studied CKDs by utilizing the large-scale omics approaches.

Paper I describes a study on the potential genetic causes of diabetic nephropathy (DN). DN is the major cause of ESRD among all CKDs worldwide. Here we studied a Finnish sibling cohort, in which sibling pairs are both affected by type 1 diabetes (T1D), but they are discordant for development of DN. Studying the genetics of DN is challenging as one is searching for genes and genomic variants that only cause disease if the patient has diabetes and hyperglycemia. The study was carried out by sequencing the whole genome of the discordant sibling pairs, and performing multiple bioinformatic analyses on the data. We studied protein altering variants and enrichment of variants in regions associated with presence or absence of DN. We replicated our findings in a larger T1D cohort of unrelated Finns with T1D, referred to as the FinnDiane cohort. We identified several top candidate genes some of which were studied in a zebrafish model. Some of the top candidate genes and genomic variants, showing highest association with the presence or absence of DN were characterized. One of them was protein kinase C epsilon that has been found to be associated with development of DN.

Paper II reports on a meta-analysis of the expression profiles of glomerular diseases. We summarized all microarray and proteomics data sets on glomerular diseases, including studies on patient biopsy and animal models. We developed a pipeline for meta-analysis on microarray data, and compared two DN human patient studies together with DN animal model studies. We have not found any consensus pathways that are significant across all glomerular diseases or disease models.

Paper III uses state-of-the-art single cell RNA sequencing technology (scRNAseq) to elucidate the expression profiles of kidney organoids. The organoids were derived from induced pluripotent human stem cells and were engineered with CRISP(e)R technology to induce fluorescent reporters facilitating the monitoring of different stages of organoid development. We observed cell clusters expressing mature podocyte and tubular markers. We also compared the transcriptomic profile of these two clusters with previous reported healthy human glomerular and tubular biopsies, and observed a similarity of organoid to adult kidney.

LIST OF SCIENTIFIC PAPERS

This thesis is based on the publications listed below.

I. **Jing Guo**, Owen J.L. Rackham, Bing He, Anne-May Österholm, Erkka Valo, Valma Harjutsalo, Carol Forsblom, Iiro Toppila, Maikki Parkkonen, Qibin Li, Wenjuan Zhu, Nathan Harmston, Sonia Chothani, Miina K. Öhman, Eudora Eng, Yang Sun, Niina Sandholm, *Enrico Petretto, *Per-Henrik Groop, *Karl Tryggvason. **Whole genome sequencing in Finnish type 1 diabetic siblings discordant for kidney disease reveals DNA variants associated with diabetic nephropathy.** *Manuscript submitted.*

II. Sam Tryggvason, **Jing Guo**, Masatoshi Nukui, Jenny Norlin, Börje Haraldsson, Hans Jörnvall, Karl Tryggvason, Liqun He. **A meta-analysis of expression signatures in glomerular disease.** *Kidney International, 2013, Volumn 84, Issue 3, Pages 591–599.*

III. Cecilia Boreström, Anna Jonebring, **Jing Guo**, Henrik Palmgren, Linda Cederblad, Anna Forslöw, Anna Svensson, Magnus Söderberg, Anna Reznichenko, Jenny Nyström, Jaakko Patrakka, Ryan Hicks, Marcello Maresca, Barbara Valastro, Anna Collén. **A CRISP(e)R view on kidney organoids allows generation of an induced pluripotent stem cell–derived kidney model for drug discovery.** *Kidney International, 2018, Epub ahead of print, PMID:30072040.*

Other publications not included in the thesis:

Sonia Zambrano, Patricia Q. Rodriguez, **Jing Guo**, Katja Möller-Hackbarth, Angelina Schwarz, and Jaakko Patrakka. **FYVE domain-containing protein ZFYVE28 regulates EGFR-signaling in podocytes but is not critical for the function of filtration barrier in mice.** *Sci Rep. 2018 Mar 16;8(1):4712. doi: 10.1038/s41598-018-23104-z.*

Xiaonan Zhang, Arjan Mofers, Per Hydbring, Maria Hägg Olofsson, **Jing Guo**, Stig Linder, and Padraig D'Arcy. **MYC is downregulated by a mitochondrial checkpoint mechanism.** *Oncotarget. 2017 Oct 6;8(52):90225-90237. doi: 10.18632/oncotarget.21653*

Bing He, Lwaki Ebarasi, Zhe Zhao, **Jing Guo**, Juha R.M. Ojala, Kjell Hultenby, Sarah De Val, Christer Betsholtz and Karl Tryggvason. **Lmx1b and FoxC Combinatorially Regulate Podocin Expression in Podocytes.** *JASN December 2014, 25 (12) 2764-2777; DOI: https://doi.org/10.1681/ASN.2012080823*

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

ACEi	Angiotensin Converting Enzyme Inhibitors
bp	base pair
ARB	Angiotensin II Receptor Blockers
CAGE	Cap Analysis of Gene Expression
cDNA	complementary Deoxyribonucleic Acid
CKD	Chronic Kidney Disease
DEG	Differentially Expressed Gene
DN	Diabetic Nephropathy
DN-RMR	Diabetic Nephropathy associated Recurrently Mutated Region
DNA	Deoxyribonucleic Acid
DSP	Discordant Sibling Pairs
eQTL	expression Quantitative Trait Loci
ERCC	External RNA Controls Consortium
ESRD	End-Stage Renal Disease
EST	Expressed Sequence Tag
FACS	Fluorescence Activated Cell Sorting
FANTOM	Functional ANnoTation Of the Mammalian genome
FDR	False Discovery Rate
FPKM/RPKM	Fragments/Reads Per Kilobase per Million mapped reads
F-SKAT	Familial Sequencing Kernel Association Test for dichotomous traits
GBM	Glomerular Basement Membrane
GFB	Glomerular Filtration Barrier
GWAS	Genome Wide Association Study
hiPSC	human induced Pluripotent Stem Cells
KEGG	Kyoto Encyclopedia of Genes and Genomes
NGS	Next Generation Sequencing
PRKCE	Protein Kinase C Epsilon
QC	Quality Control
RMR	Recurrently Mutated Region
RNA	Ribonucleic Acid

RNA-Seq	RNA Sequencing
SAGE	Serial Analysis of Gene Expression
scRNA-Seq	single cell RNA Sequencing
SKAT	Sequencing Kernel Association Test
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variant
T1D	Type 1 Diabetes
T2D	Type 2 Diabetes
TPM	Transcripts Per Kilobase per Million mapped reads
tSNE	t-distributed Stochastic Neighbor Embedding
UMI	Unique Molecular Identifiers
WGS	Whole Genome Sequencing

1 BACKGROUND

1.1 CHRONIC KIDNEY DISEASE

1.1.1 Kidney and glomeruli

The primary function of the kidney is to filter out the small molecular waste products, excess water and electrolytes from blood in order to maintain homeostasis in the body. Besides generation of urine, kidneys have also an important role in production of hormones, such as erythropoietin and renin.

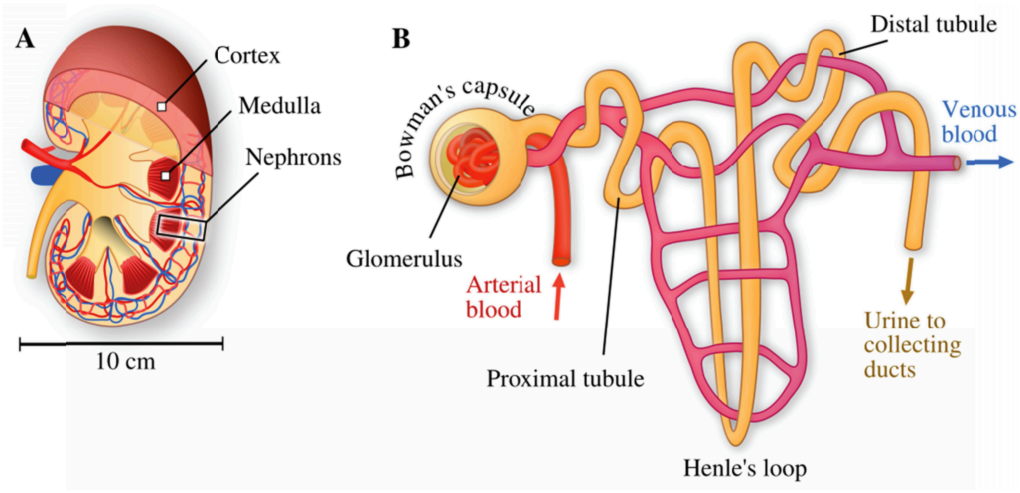

Figure 1: A: Principal structure of the human kidney. The nephron is the basic functional unit. B: The structure of a nephron. Blood enters glomeruli where water and small molecules are filtered out as the primary urine. The majority of in primary urine is reabsorbed in the tubules and returned to blood circulation. Figure modified from (Mäkinen 2010).

Nephron is the basic functional unit of the kidney (**Figure 1**). There are around 0.8-1.5 million nephrons in an adult human kidney. Each nephron consists of a proximal end, the glomerular tuft, located in the Bowman's capsule, and a long renal tubule that extends through different segments of the kidney ending with the collecting duct that opens into the medulla. The kidney receives about 25 % of the cardiac output and the glomerulus is responsible for the ultrafiltration of blood, generating about 180 liters of primary urine a day.

About 99% of the primary urine is reabsorbed in the tubuli so that and eventually only about a total of 1-1.5 litres of concentrated urine is excreted from the body daily. The ultrafiltration of blood occurs in the capillary wall of the glomerulus which contains of three layers: fenestrated glomerular endothelial cells, the glomerular basement membrane (GBM), and podocyte foot processes located on the outer surface of the glomerular capillary (**Figure 2**). This glomerular filtration barrier (GFB) prevents large molecules such as proteins like albumin and blood cells from entering the urine, while allowing water, electrolytes and other small molecules to be retained. Mesangial cell is the third cell type of the glomerulus. They are thought to act as pericytes of the glomerulus.

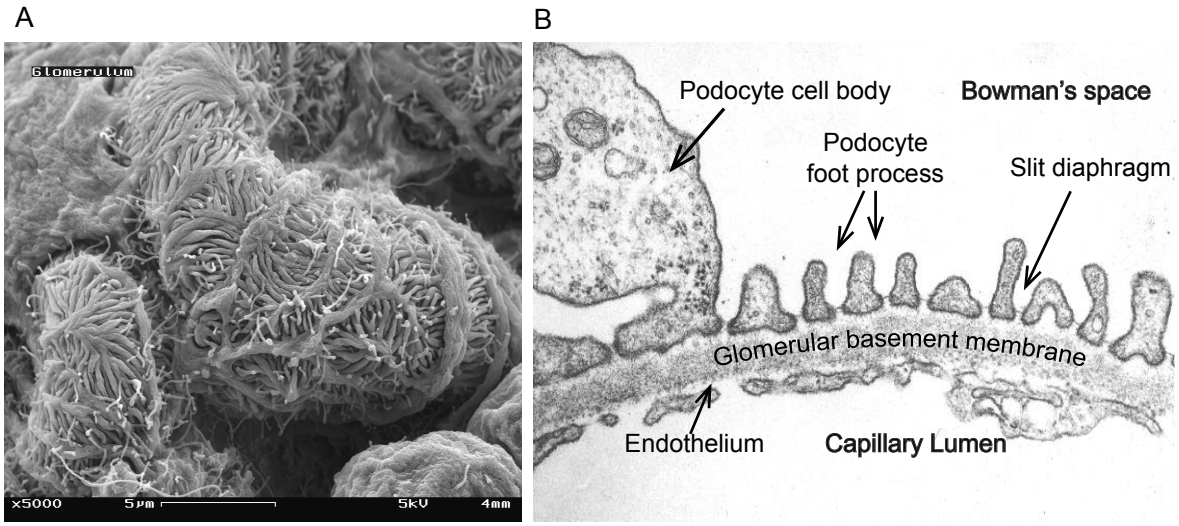

Figure 2. A. Electron microscopy scanning of a mouse glomerulus. The outer surfaced of the capillary is covered by food processes of podocytes, such that foot processes of two neighboring podocytes for interdigitating processes. Magnification x5,000. Image source: wikidoc.org. B. Cross section of the glomerular filtration barrier. The endothelial cells form a single layer on the inside of the capillary with numerous fenestrae that are devoid of a physical membrane. The glomerular basement membrane (GBM) is located between the endothelial cells and the podocyte food processes and has uniform thickness of about 300-350 nm in humans. The outside of the capillary is covered by podocyte foot processes that are partially embedded in the GBM. The food processes are separated by an uniformly wide slit diaphragm that forms a physical hinder for proteins larger than albumin (69 kDa). Image adapted from educational slide from New York Uni. Langone Medical Center.

1.1.2 Glomerular disorders are a major medical challenge

Glomerular disease processes are responsible for >70% of end-stage renal disease cases (ESRD). These include complications of systemic disorders, such as diabetes and hypertension, but also many primary glomerular diseases, such as IgA nephropathy and membranous nephropathy. In most glomerular disease processes, the ultrafiltration barrier fails which results in leakage of albumin to urine, albuminuria. Albuminuria is the hallmark sign of most renal diseases.

Glomerular diseases often show similar histopathological features, which include mesangial matrix expansion, mesangial cell proliferation, podocyte foot process effacement and eventually podocyte loss. The molecular mechanisms driving the progression of these changes are still poorly understood. This is mainly due to the fact that there is a very limited access to diseased human glomerular samples. Also, poor translation in kidney diseases from animal models to man contributes to our poor understanding of pathobiology in human glomerulopathies.

1.1.3 Diabetic nephropathy

Diabetic nephropathy (DN) is the single leading cause of ESRD worldwide (Jha *et al.* 2013), takes up to 50% cases in developed world. It is the main cause of morbidity and mortality in

diabetes patients (Orchard *et al.* 2010, Collins *et al.* 2012, Tuttle *et al.* 2014). The number of patients with type 2 diabetes (T2D) was estimated to be 415 million worldwide in 2015 and is predicted to affect 642 million people by 2040 (IDF 2015), and approximately 30% will eventually develop DN (Afkarian *et al.* 2016). This morbidity of DN is even higher among type 1 diabetic (T1D) patients, which is estimated to be up to 40% (Gubitosi-Klug *et al.* 2008, Gheith *et al.* 2016).

The major histological changes of DN are detected in the glomeruli, i.e. mesangial expansion, thickening of the GBM, foot process effacement and at later stages glomerular sclerosis (**Figure 3**). Clinically, the cardinal sign of DN, i.e. leakage of albumin to the urine is measured by albuminuria excretion rate (μg/min). If albumin leakage is minimal (30-299 μg/min), it is referred to as microalbuminuria and when in more substantial amount (>300 μg/min), it is referred to as macroalbuminuria. In many cases of macroalbuminuria, DN leads to ESRD, a condition treatable only with chronic dialysis or kidney transplantation. Sadly, 70% of ESRD patients die within 5 years on dialysis (O'Shaughnessy *et al.* 2015).

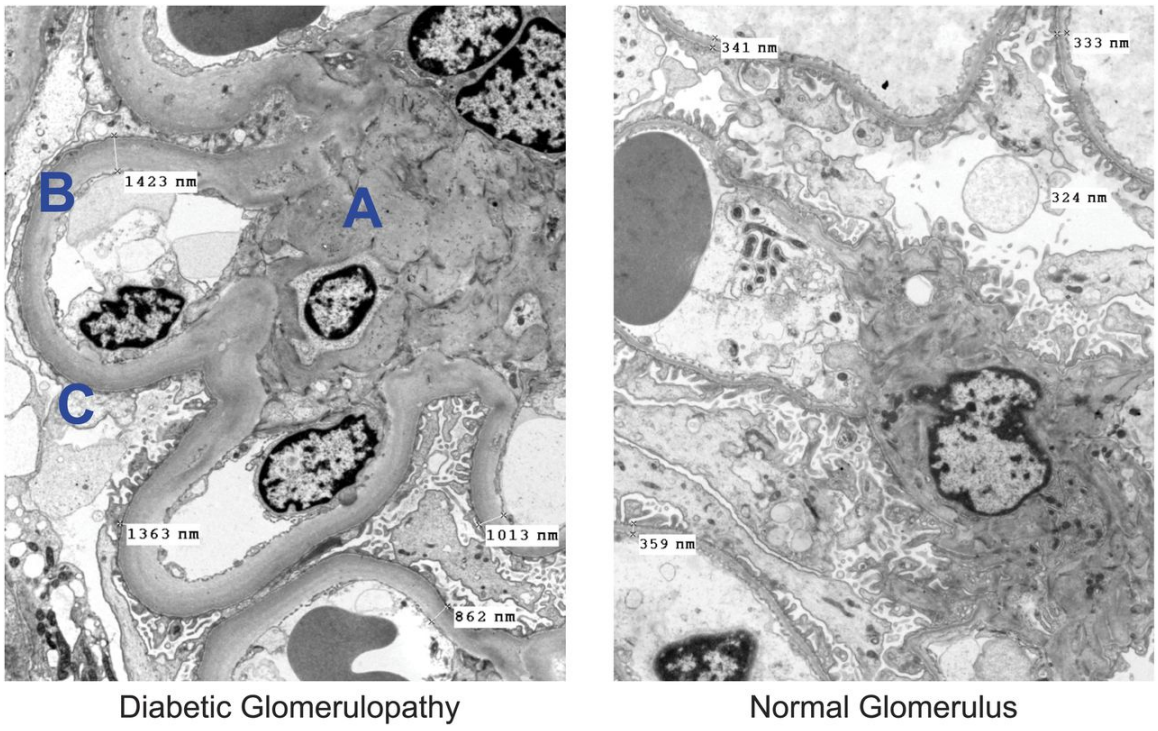

Figure 3. Structural changes of glomerulus in diabetic nephropathy (DN) by electron microscope, magnification x3500. Comparing to normal glomeruli, the typical features in DN glomeruli are: A. mesangial expansion; B. glomerular basement membrane thickening; C. food process enfacement. Images from (Alicic *et al.* 2017).

Multiple risk factors contribute to DN. A key factor is persistent hyperglycemia. It has been shown that strict glycemic control can significantly reduce the occurrence of DN (Reichard *et al.* 1993). However, despite the significant improvement of treatment for diabetes mellitus over 30 years, the occurrence of DN is not substantially decreased (Gregg *et al.* 2014). It is reported that even under the strictest control of blood glucose the cumulative incidence of DN remains 9% after 30 years of T1D (Nathan *et al.* 2009). How does hyperglycemia lead to DN? This is a very complex question, which is still a major challenge to answer. There are

many potential mechanisms, which involves several pathways. One basic mechanism is that cells fail to down-regulate their glucose transporters, which results in hyperglycemic conditions intracellularly. This, in turn, results in a metabolic dysregulation that causes for instance dysfunction of mitochondria leading to increased production of reactive oxygen species and superoxide that drive the development of cellular damage. Extracellularly, hyperglycemia can cause for instance non-enzymatic glycosylation of target molecules, which can contribute to the development of tissue damage.

Other modifiable factors are hypertension, dyslipidemia and smoking, age, race, and genetic profiles (Lim 2014). It has also been reported that male gender is a risk factor for ESRD in T1D patients (Harjutsalo *et al.* 2011), while female gender is protective from ESRD (Iseki *et al.* 1996). However, The gender effect on DN remains controversial as gender effects can also be confounded by age, ethnic group, or other factors (Iseki 2008), and gender variation is often absent in a large cohort of chronic kidney disease (CKD) (Silbiger *et al.* 2008). Genetic factor in DN is discussed in another section below.

Treatment to prevent the progression of DN is very limited. The most effective ones are angiotensin converting enzyme inhibitors (ACEi) and angiotensin II receptor blockers (ARB), which are functional by reducing the intra-glomerular pressure (Tuttle *et al.* 2014). They can significantly alleviate the deterioration together with hypertension treatment in clinic (Lim 2014). Recent studies have also shown that sodium glucose linked transporter 2 (SGLT2) inhibitors can slow down the progression and lower the rate of clinical relevant renal events (Wanner *et al.* 2016, Neal *et al.* 2017). The mechanism is to prevent the glucose reabsorption in proximal tubuli thus change the hemodynamics in kidney (Vallon *et al.* 2017). However, those drugs cannot reverse the progression, and cannot address the primary disease mechanisms that are still largely unknown.

1.2 GENETICS

1.2.1 Human genetics

The genetic information carried in genomic DNA (Deoxyribonucleic acid) is the blueprint for most living organisms. DNA is essentially the sequence of four types of nucleotides: Adenine (A), thymine (T), cytosine (C) and guanine (G). It forms a double helix structure that is composed of two complementary strands, i.e. A is complementary to T and C to G. Each of the complementary pairs is called base pairs.

Homo Sapiens, i.e. we human, are diploid organisms. Our genome is packed within 22 pairs of autosomal chromosomes and one pair of sex chromosomes (XX for female and XY for male). The homologous pair of chromosomes contains hereditary information from mother and father. Nucleotides at a locus of a chromosome can be identical (homozygous) or different (heterozygous).

The Human Genome Project and a parallel project taken by Celera Corporation published the first complete sequence of human genome in 2001 (Lander *et al.* 2001, Venter *et al.* 2001). It

provides the first draft of human genome and has greatly benefited the genetic study in human diseases.

The current human reference genome contains 20,418 coding genes, 22,107 non-coding and 15,195 pseudo-genes [Ensembl (Zerbino *et al.* 2018), gene build GRCh38.p12]. Only 1.2% of the genome is protein coding, and 24% is introns or segments upstream/downstream of the coding regions. The remaining genome is intergenic and its role and functions are still poorly known. Around 99.5% of the human genome is identical between two individuals, regardless of ethnical or phenotypical appearance of the individual.

1.2.2 Genotyping methods - a brief introduction

The DNA sequencing was first developed by Sanger Frederick (Sanger *et al.* 1975). It was based on DNA replication using specific chain-terminating dideoxynucleotides. Read length from Sanger sequencing can reach 800 base pairs of nucleotides. The method was the predominant technology for DNA sequencing until late 2000s. It has greatly benefit the progress of genetic studies.

Next generation sequencing (NGS), "next generation" to the Sanger sequencing technology, is a massive parallel DNA sequencing technology. DNA sequences are fragmented into short reads, and amplified by polymerase chain reaction (PCR). Adapters that could bind on the sequencing chip are added to the end of DNA sequences. The reads are sequenced by commercially available sequencing machines, such as Illumina HiSeq and Complete Genomics (the later BGISEQ). The fragmented reads, typically 30-300 bp long, can then be assembled *de novo* or by reference genome. With intensive and collaborative work, it took a decade to complete sequencing of the first human genome. Nowadays it only takes one day to sequence a complete human genome by NGS, which is referred as whole genome sequencing (WGS).

Besides sequencing methods, SNP arrays using hybridization principal are widely used for large-scale SNP detection. Hybridization of DNA and probe sequences is more favorable when they are fully complementary to each other, and less favorable for mismatching SNP sites. Thus the intensity signal of hybridization can be used to detect the genotyping of the SNP. This method has been widely used for genome-wide association study because of it's high-throughput of SNP detection, and efficiency both economic-wise and labour-wise. However, with the declining cost of NGS, it is slowly overtaken by WGS.

1.2.3 Genetic studies in DN

A genetic risk factor of DN was first reported in a study of Finnish families in 1989 (Seaquist *et al.* 1989). In families where a proband had ESRD, 24 out of 26 diabetic siblings (83%) was detected to have DN, whereas in families where the proband did not have DN; only 2 out of 11 (17%) diabetic siblings had DN. A similar conclusion has been observed in a study on Danish T1D families (Borch-Johnsen *et al.* 1992). A later epidemiological study in Finnish

T1D patients suggested that with one sibling affected by DN, the risk of other T1D siblings being affected by kidney disease was increased over two-fold (Harjutsalo *et al.* 2004).

To study the genetic factors of a disease or trait, there are two main approaches. The traditional linkage study uses the family pedigree data to identify linkage of alleles at a genetic susceptible locus and know genetic marker locus through generations of families. Genome-wide markers are tested in pedigrees segregating a trait. The second approach is genome-wide association study (GWAS) to identify associations between a locus and phenotypical trait. It involves collection of genetic information of unrelated, affected (cases) and unaffected (controls) individuals, mostly by SNP arrays, which allow genotyping of millions of SNPs on one single chip. This approach has gained extensive popularity with the declined cost of genotyping chips, and larger international cooperation.

Linkage studies based on sibling pairs – either with discordant sib-pairs (DSP) where both are affected by diabetes but discordant for DN, or with "affected" sib-pairs where both are affected by DN – suggests a lineage peak on chromosome (Imperatore *et al.* 1998, Moczulski *et al.* 1998, Österholm *et al.* 2007, He *et al.* 2009). The limitation of the traditional linkage approach is that the tested significant region is usually mega base pair long, and may need further fine-mapping and functional annotation. Furthermore, as the study object is familial data, the results shall always be tested and replicated in other families and populations.

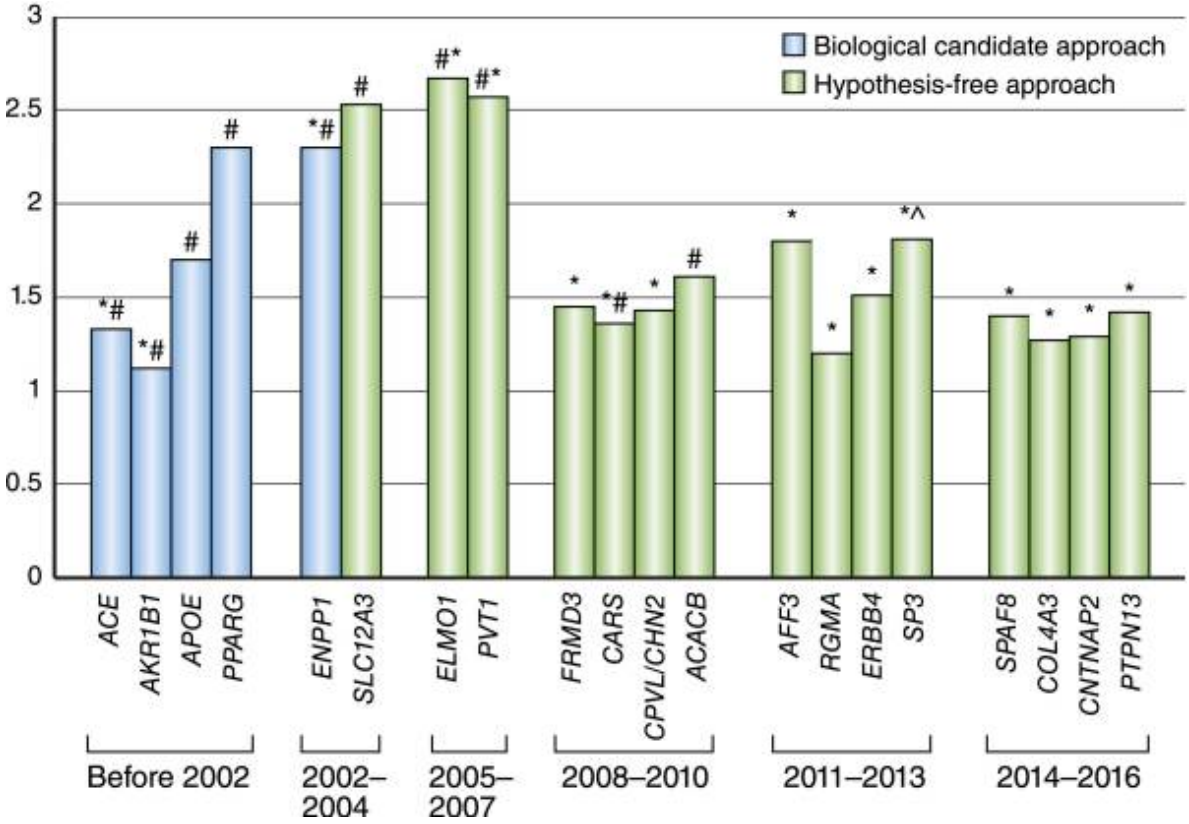

Figure 4. Progress of identification of diabetic nephropathy (DN) associated genes by genome-wide association study (GWAS). Plot derived from (Ma *et al.* 2017) and gene list derived from (Ma 2016). *Discovered/replicated in studies in subjects with T1D. #Discovered/replicated in studies in subjects with T2D. ^Evidence of sex difference in the association signal, with significant association detected only for women. Most variants listed in the plot does not reach genome-wide significance threshold.

GWAS, benefiting from a large-scale test of SNPs, and large sample size, commonly leads to very high significant statistics. Although it has lead to fruitful discovery for other diseases (Visscher *et al.* 2017), there are only a few genes been associated to DN as summarized in **Figure 4**.

One limitation of applying GWAS on DN is the insufficient sample size. Although diabetes almost has epidemic proportions in the modern society, the majority of patients with T2D has late onset of diabetes in their life course, and it usually takes another 10-20 years to develop kidney complication (Gheith *et al.* 2016). The samples size of DN affected individual can hardly reach level of ten thousands of samples, and significance of variants identified can barely reach 10^{-8}. Additionally, clinical diagnoses can be complicated for DN (Roshan *et al.* 2013), diabetic patients with nephropathy have different histopathological features, and may caused by other reasons (Alsaad *et al.* 2007). While the relationship of statistical significance and actual biological function is another debate, the power of GWAS approach in case of DN is limited.

A more recent study by FinnDiane performed GWAS and replication in 12,540 Finnish T1D patients and sequenced the whole exome of 997 patients (Sandholm *et al.* 2017). There were no single variants reached genome significant level in the study, despite the enlarged sample size. However, with joint meta-analysis of two stages of clinical diagnoses, three variants showed suggestive association with DN. And association of genes with specific DN related pathways are reported, suggesting a careful phenotypic classification for GWAS trait is necessary for detection of meaning biological signals.

With easy access to WGS technology, it is possible now to detect more low-frequency variants with intermediate impact (Cirulli *et al.* 2010). The decline of expenses for WGS allows the application on larger sample size, though it is still not possible to reach as much as for GWAS. While GWAS is detecting signals spreading across the genome (depending on the array), suggesting possible associations for certain area, WGS is more likely to give a true signal.

1.3 TRANSCRIPTOMICS STUDY: HISTORY, PRESENT AND FUTURE

1.3.1 Introduction to transcriptome

Transcriptome represents the whole RNA (Ribonucleic Acid) transcripts in an organism. RNA is synthesized using genomic DNA as template, and then directs the expression of protein. This process is referred as central dogma (Crick 1958, Crick 1970) in cell biology. Besides the role in the classic DNA-RNA-protein, recent studies also show other important roles of RNA (Cech *et al.* 2014).

A transcriptomics study examines the expression levels of RNA in a cell, tissue or organism. It captures a snapshot of RNA molecules in a designed condition. It can be used to examine particular types of RNA, or the total RNA. The messenger RNA (mRNA) is most widely studied in transcriptomics as it directly reflects the gene expression. mRNA carries the

information from DNA and its sequence is translated to amino acid sequence, which is then assembled into protein. Other types, so called non-coding RNA, are important in translation and regulation of cell functions. For instance, ribosomal RNAs (nRNA) and transfer RNAs (tRNA) are important in mRNA translation. Another class, long non-coding RNA (lncRNA), which is an artificially defined (>200 nucleotides) to distinguish from other small RNA types, has been associated to several diseases (Wilusz *et al.* 2009, Wapinski *et al.* 2011). Projects like FANTOM (Functional ANnoTation Of the Mammalian genome) (Lizio *et al.* 2015) and GENCODE (Harrow *et al.* 2012) are taking large collaborative efforts to study the function of lncRNA (Derrien *et al.* 2012, Hon *et al.* 2017).

1.3.2 Evolution of methods for transcriptomics study

The study of transcriptome is greatly benefited by the rapid development of biotechnology. For individual or small sets of transcripts, the quantification can be measured using northern blotting, and quantitative reverse transcription polymerase chain reaction (RT-qPCR). The later is still widely in us, and is a tool for validation for large-scale transcriptomics study. For a larger set of transcripts, libraries of mRNA can be collected and preserved by converting instable mRNA into its complementary DNA (cDNA) back in the 1970s (Sim *et al.* 1979). The short sub-sequences of cDNA are called expressed sequence tags (ESTs) (Adams *et al.* 1991). The quantity of cDNA has a representation of the amount of its complementary transcripts. These libraries of individual transcripts can be sequenced by methods like Sanger sequencing (Sanger *et al.* 1975). It greatly benefits the discovery of transcripts in an organism before we could capture the whole picture of its transcriptome.

Based on EST, a higher throughput method, serial analysis of gene expression (SAGE) (Velculescu *et al.* 1995), was developed in 1990s to identify and quantify thousands of transcripts at one time. EST libraries are digested into 11bp “tags” by restriction enzymes, and then concatenated into >500 bp long strands. The long strands of concatenated cDNA are then sequenced by Sanger sequencing. Cap analysis of gene expression (CAGE) method (Shiraki *et al.* 2003) applies the same methods but sequences only the transcriptional start site of genes. CAGE is still used to the promoter analysis by projects like FANTOM to date. SAGE and CAGE have limitations of intensive labour work and high cost, and were largely overtook by microarray and deep sequencing or so called RNA sequencing (RNA-seq) methods in the early 2000s.

1.3.3 Microarray and RNA-seq

The most widely used technologies in transcriptomics are microarray and RNA-seq. Microarray technology was first introduced in 1995 (Schena *et al.* 1995, Pozhitkov *et al.* 2007) and remained its popularity until the early 2010s. Microarray requires prior transcript information on the organism to be studied. Short nucleotide oligomers called “probes” with complementary sequences of organism transcripts are pre-installed on a physical slide made by glass or silicon. cDNAs that have been labeled with fluorescence are hybridized to the probes. The transcript abundance can then be determined by the intensity of the fluorescence

at each probe sets (Barbulovic-Nad *et al.* 2006). Commercial high-density arrays like Affymetrix GeneChip array (Santa Clara, California) and Illumina BeadChip array are available for popularly studied organisms. These microarray slides allow hybridization of ten thousands probe sets on one single slide, and lower down the cost and labour work dramatically. Choices of microarrays are available to detect different types of RNA molecules. Microarray was the predominant tool for transcriptomics study from early 2000s to mid 2010s. Therefore the transcriptomics analyses in **Paper II** (published in 2013) are largely based on this technology.

Commercial microarray platform provide standardized software and algorithm to normalize the image intensity produced from microarray chips. The raw data needs to be normalized to adjust for the background noise of the image, labeling efficiency and laser detection, etc. The most used normalization methods are robust multi-array average method (RMA) (Irizarry *et al.* 2003) and GC-RMA (Wu *et al.* 2004) for Affymetrix platform. The normalization corrects the raw intensity data by background correction, quantile normalization, gene-wise correction, and a logarithm transformation of expression value. Then cross sample quality control (QC) like MA plots, RNA degradation plot can be applied on a data set to estimate the quality of microarray data. Data after QC can then be used for further analysis, like differential expression analysis, Gene Set Enrichment Analysis (GSEA), Over-Representation Analysis (ORA), etc.

RNA-Seq sequences the transcript cDNA in depth using NGS technology (Morozova *et al.* 2009, Wang *et al.* 2009, Ozsolak *et al.* 2011). **Figure 5** shows the principal procedure for a RNA-seq experiment. RNA extracted from tissue or cells can be selected by different protocol to isolate a specific type of RNA. For example, mRNA is enriched by poly-A selection in this step. Long transcripts are fragmented into 200-300 bp, and reverse transcribed into cDNA. Adapters designed for sequencing platforms are added to single or both ends of the cDNA fragments. After PCR amplification, library is constructed and ready to be sequenced. In the sequencing procedure, the sequencer detects the nucleotides in cDNA fragments, and generates short reads (30-400bp depending on platforms) of single-end or pair-end.

A standard bioinformatics analysis for RNA-seq has following steps: First, the sequencing reads are aligned to reference genome by tools such as STAR (Dobin *et al.* 2013), Tophat2 (Kim *et al.* 2013). The aligned sequences are outputted in Binary Alignment/Map (BAM) format. Secondly, assemble transcripts transcripts using a reference transcript annotation, *de novo*, or combined approach (Martin *et al.* 2011). Thirdly, Number of reads within exons of a gene or transcript is counted by tools like HTSeq (Anders *et al.* 2015) or featureCounts (Liao *et al.* 2014). Counts are normalized upon the factors like sequencing library size, gene length, CG content, etc. Reads/Fragments per kilobase per million mapped reads (RPKM/FPKM) was used at the first studies of RNA-seq. The method normalizes counts based on library size and gene length. The later introduced transcripts per kilobase per million mapped reads (TPM) is similar to RPKM, but becoming more popular as it results a equal number of TPMs

within each sample, thus more suitable for comparison between samples. Then the normalized counts can be used for downstream analysis similarly to microarray.

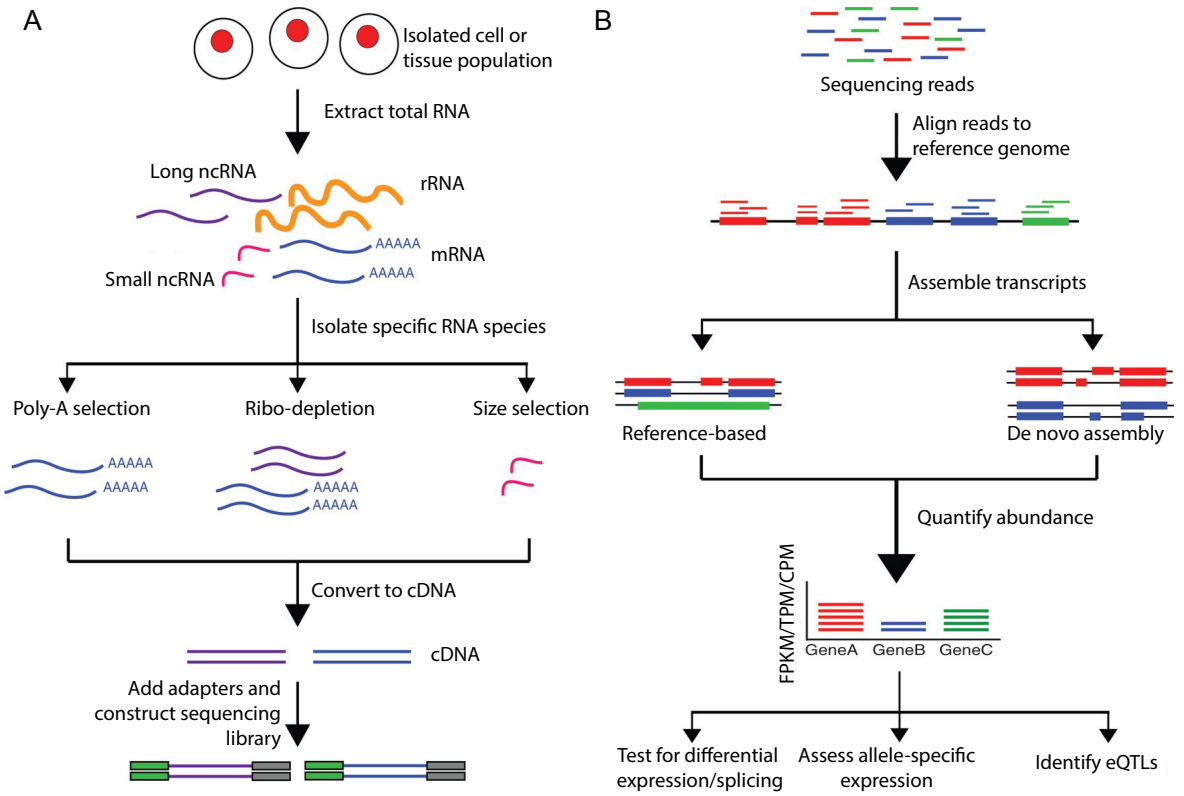

Figure 5. RNA sequencing (RNA-seq) workflow. A. Sequencing library construction: first extract the total RNA from biological material of choice. Subtract type(s) of RNA can be isolated using specific protocols. For example, poly-A selection can enrich polyadenylated mRNA, ribo-depletion protocol removes ribosomal RNAs, or size selection process to keep only small RNAs. RNA is then reverse-transcribed to complementary DNA (cDNA) and sequencing adapters are added to the end of cDNA fragments. After amplification by PCR, the library is by ready to be sequenced. B. Principal bioinformatic analyses for RNA-seq. Sequencing reads are aligned to reference genome. Transcripts can be assembled using reference transcript annotation, or *de novo* approach to identify novel transcripts. The expression level of each gene is estimated by normalizing counts that aligned to the transcript region. Then downstream analyses like differential expression can be applied. Figure adapted from (Kukurba *et al.* 2015).

As shown in **Figure 6**, RNA-seq has overtook microarray as the first choice for transcriptomics study (Su *et al.* 2014). RNA-seq has desired advantages: it does not require previous knowledge of the organism for probe design, and it is not affected by SNVs in the probe sequence. The detection of transcripts in RNA-seq is not limited by probe sets, thus could include splicing variants, allele specific expression and small RNAs (Nookaew *et al.* 2012). Studies have shown a high consistency between microarray and RNA-seq experiment (Nookaew *et al.* 2012, Zhao *et al.* 2014). However, RNA-seq has better dynamic range for highly and lowly expressed genes, lower technical variant and higher accuracy for expression level (Hoen *et al.* 2008, Wang *et al.* 2009, Nookaew *et al.* 2012). Moreover, recent technology has made it possible to sequence the whole transcriptome from as little material as of single cell.

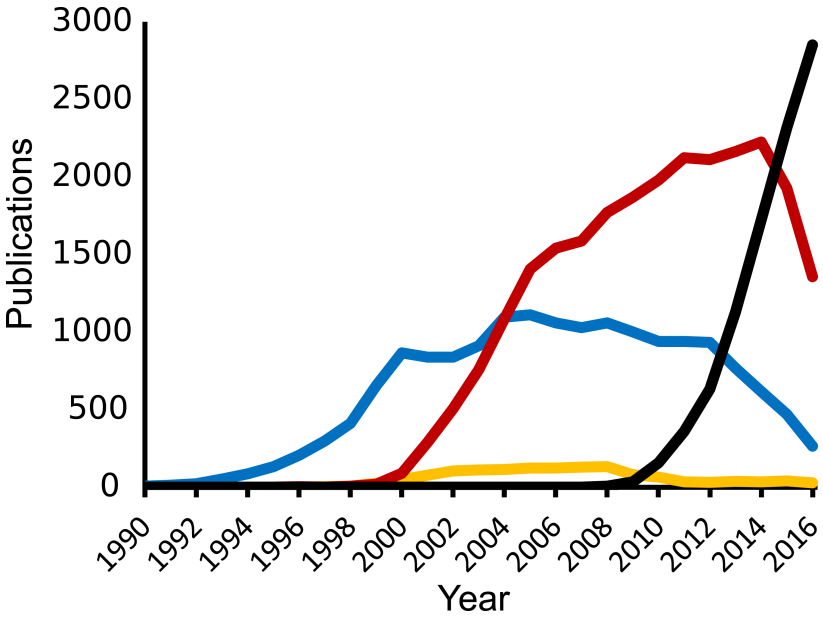

Figure 6. Number of publication of transcriptomics study using expressed sequence tag (EST, blue), RNA microarray (red), RNA sequencing (RNA-seq, black), and serial/cap analysis of gene expression (SAGE/CAGE, yellow). RNA-seq technology has become the predominant tool for transcriptomics study. Plot derived from (Lowe *et al.* 2017).

1.3.4 New promising approach: single cell sequencing

Transcriptomics studies in kidney field has yielded to discovery of new biomarkers of CKD (Mishra *et al.* 2003, Ju *et al.* 2015) and potentially beneficial for novel therapies (Kretzler *et al.* 2018). However, there is still limited knowledge about expression patterns and their changes at the level of individual cells and cell types. Kidney has a complex structure with at least 18 distinct cell types (Kretzler *et al.* 2018). The resolution of transcriptome on single cell level is expected to benefit greatly the researches in kidney.

1.3.4.1 scRNA-seq methods

The first mRNA sequencing at the single cell level was reported by (Tang *et al.* 2009). Subsequently, protocols were developed and improved by various groups and applied for detailed analyses of whole tissues or specific cell types. A typical protocol for scRNA-seq library construction includes organ or tissue dissociation, single cell capture, cell lysis, revers transcription, and amplification. Then the library can be sequenced as bulk RNA-seq. The major differences between protocols are the methods for single cell capture (**Figure 7**), reverse transcription and amplification.

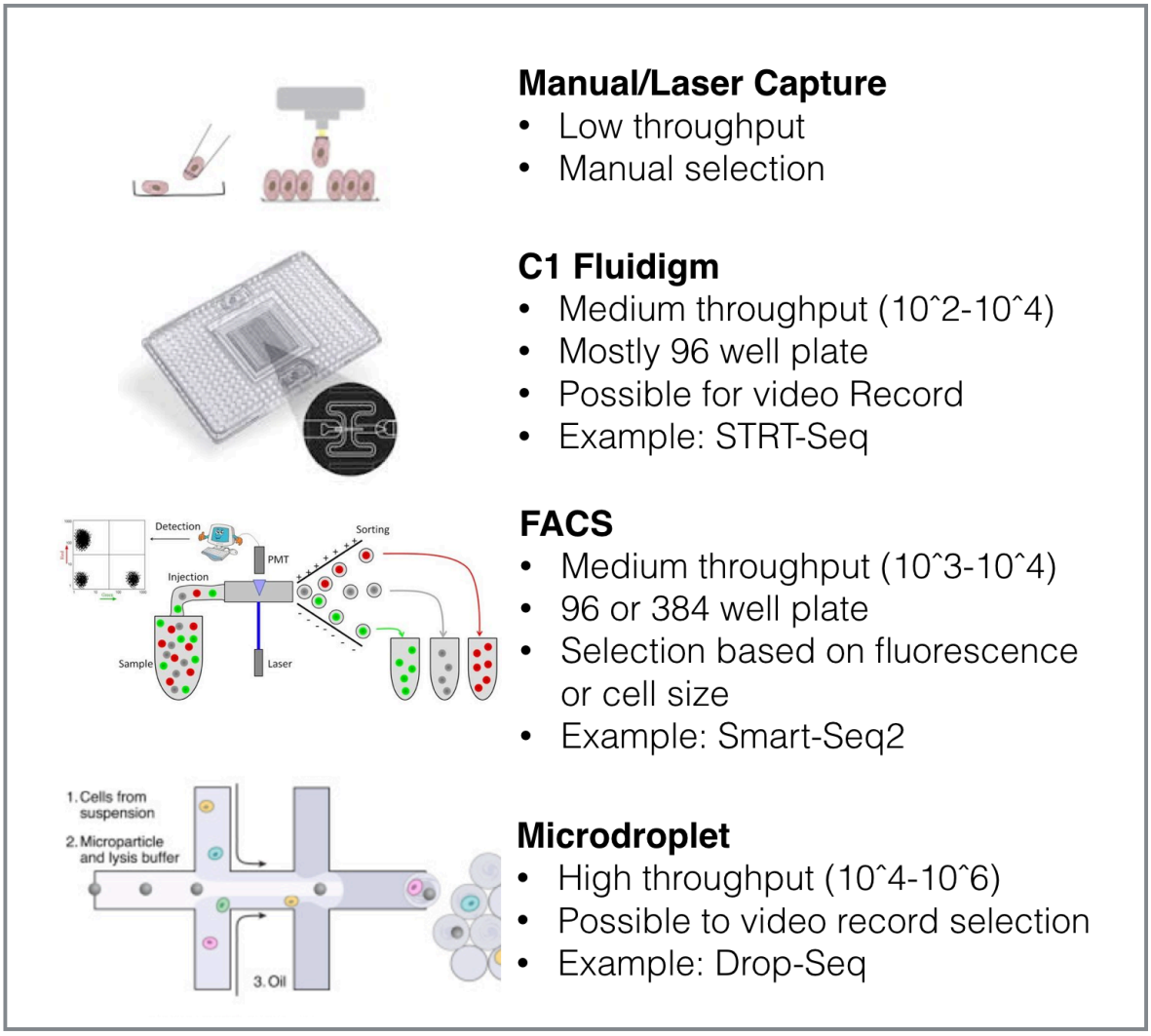

Figure 7. Different methods used for single cell capture and their features. Adapted from (Kolodziejczyk *et al.* 2015).

For the requirement of higher throughput per experiment, the method based on micro droplet (Macosko *et al.* 2015) is gaining more popularity. The commercially available platform 10x Genomics is based on this method. The first publications use reverse transcription method by polyA tailing plus second strand synthesis (Tang *et al.* 2009, Tang *et al.* 2010). However, the other single-cell tagged reverse transcription (STRT) (Islam *et al.* 2011) is later more widely in use as it has higher efficiency for transcript capture. PCR amplification is widely used to enrich cDNA library, however, other method like IVT (*in vivo* transcription) (Luo *et al.* 1999) using a linear amplification is also applied on methods such as CEL-seq (Macosko *et al.* 2015) and MARS-seq (Jaitin *et al.* 2014). Both amplification methods may introduce bias, e.g. CG bias for PCR, and 3'bias for IVT.

Method used by **Paper III** is based on SMART-seq2 protocol (Picelli *et al.* 2014). The library construction procedure includes: cell dissociation, fluorescence activated cell sorting (FACS) into 384 plate, cell lysis, STRT reverse transcription, adapter ligation, PCR amplification and sequencing. SMART-seq2 has the advantage of high RNA capture efficiency(Ziegenhain *et al.* 2017), which allows a deeper sequencing of transcriptome. The

sequencing reads can cover whole transcript, instead of only the 3' end of polyA RNA, thus it provides opportunity to detect splicing events.

1.3.4.2 Bioinformatics for scRNA-seq analysis

The design of a single cell transcriptome analysis shall be based on a careful examination of methods and research questions. It is crucial to have a good balance of cell number and biological complexity (Grun *et al.* 2015). Other techniques can also be considered in the experimental design. For example, the usage of Unique Molecular Identifiers (UMI) can significantly reduce amplification bias (Kivioja *et al.* 2011), and external RNA control spike-ins (Baker *et al.* 2005) such as ERCC (External RNA Controls Consortium) be used for quality control and count normalization.

Analysis of single cell transcriptomics is generally guided by bulk transcriptomic analysis. A standard tool like fastQC can be used for quality control, and mapping tools like Tophat2 and STAR are used in practice. Quantification based on isoform, i.e. Cufflinks (Trapnell *et al.* 2012) is not ideal due to the low coverage of the transcriptome. Alternatively mapping merged to genes can maximize the usage of reads and reduce ambiguity.

Ideally, different cell types in a heterogeneous tissue can be separated by unbiased cluster based on the expression profiling of each cell. It is reported that clear visual separation of cell subgroups can be obtained by using only first two principle components in a PCA scatterplot (Pollen *et al.* 2014, Zeisel *et al.* 2015). More sophisticated models (Jaitin *et al.* 2014) can be used to minimize the technical variability. T-distributed Stochastic Neighbor Embedding (tSNE) (Maaten *et al.* 2008) is commonly used to visualize the cell clustering in 2 dimension plot. Methods like SC3 (Kiselev *et al.* 2017), Seurat (Satija *et al.* 2015) are popular for cell classification.

1.3.4.3 Current scRNA-seq study in kidney

Applications of scRNA-seq in kidney field started to boom from 2017. Several recent studies with large cell number have shed new light on kidney research. Susztak's lab ((Park *et al.* 2018) sequenced 57,979 cells from adult mouse kidney, and reported a comprehensive single cell atlas of mouse kidney. (Menon *et al.* 2018) sequenced 6,414 cells from five speciments of human fetal kidney and reportd 11 clusters of specifc renal cell types. They further revealed the subclustering of progenitor, intermediate and mature stage of renal cell lineage during development. (Young *et al.* 2018) sequenced 72,501 cells from human renal tumor cells and healthy fetal, pediatric, adult cells, and marks the first scRNA-seq study on renal carcinoma. (Wu *et al.* 2018) reported a scRNA-seq study comparing human allograft kidney specimen to healthy kidney, and identified immune responses in different cell types. (Adam *et al.* 2017) presented a method to improve the dissociaiton of single cells using a protease which have high activity in the cold. The method could protentially help with the efficiency of single cell dissociation by minimizing the RNA degradation. Most of the studies made the raw data and resources publicly available, which will faciliate the further validation and meta-analysis.

2 AIMS OF THE THESIS

The overall aim of the thesis project is to generate new knowledge on the kidney and how gene expression, genomic variants can vary in diseases, particularly DN, using the latest methodologies of genomics and expression analyses, WGS and scRNA-seq.

The specific aims are:

- to identify genetic varianats in diabetic nephropathy in a Finnish T1D sib pairs cohort and attempt to identify specific mutations and variants that may predispose to DN (Paper I)
- to find common markers or pathways for glomerulus damage in a meta-analysis study (Paper II)
- to study features of a kidney organoid model derived from induced human pluripotent stem cells using scRNA-seq technology (Paper III)

3 PRESENT INVESTIGATION AND DISCUSSION

3.1 PAPER I: GENETIC ARCHITECHTURE OF DIABETIC NEPHROPATHY IN A FINNISH T1D COHORT

3.1.1 Specificity of study cohort

Genetic predisposition plays a crucial role in DN. We designed our study using a highly defined case-control cohort of discordant sibling pairs (DSP), and collected genomic information from whole genome sequencing technology. The DSPs are siblings who are both affected by type I diabetes, and cases are the sibs with kidney complications and controls have not developed DN during over 17-35 years of T1D. In this way, we get the benefit of familial study, and scan through the whole genome including rare variants. We sequenced 76 DSPs plus 3 families with 3 discordant siblings (**Figure 8a**). Moreover, we verified the results from a T1D cohort with over 4000 unrelated cases and controls in the Finnish T1D cohort FinnDiane (Sandholm *et al.* 2017).

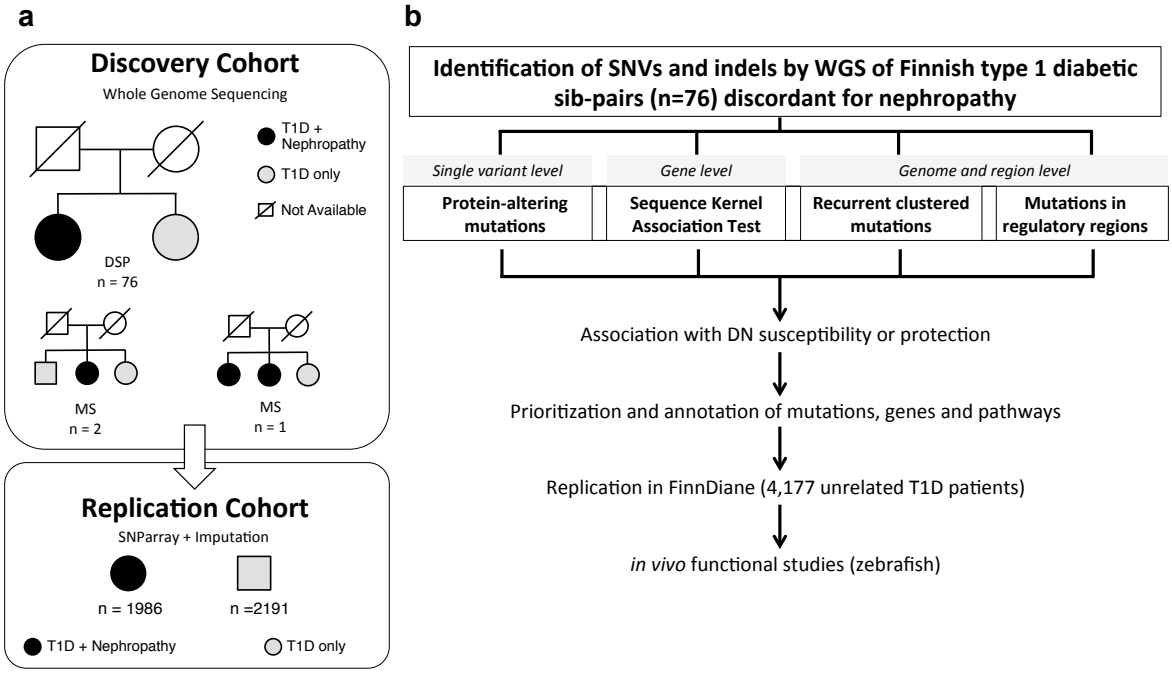

Figure 8. Discovery and replication cohorts and study design. (a) The discovery cohorts used in the search for DN susceptibility genes in Finnish type 1 diabetes (T1D) patients: the genomes of a total of 76 sib pairs concordant for T1D but discordant for diabetic nephropathy (DSPs) were subjected to whole genome sequencing (WGS). Additionally, T1D siblings from three families with 3 siblings (Multiple Siblings, MS) with or without diabetic nephropathy (DN) were included in the sequencing analyses. The control siblings (81) have had diabetes for at least 15 years [range 15-37] without developing DN, and have never been on ACE-I or ARB medication for kidney disease. The case siblings (80) have had overt proteinuria, been on dialysis, received a kidney transplant or have died from kidney complication. (b) Multi-level strategy used to analyse the WGS data from Finnish T1D individuals with or without diabetes complications. Figure derived from Paper I.

3.1.2 Genetic tests on DSPs

As a result of WGS, we detected >12 millions SNVs. To identify the potential causal and protective variants, we performed three levels of analysis (**Figure 8b**):

1. **Single variant level**: Direct comparison of cases and controls to identify those variants that are only detected in cases and only in controls in the sibling cohort. The variants are then further validated in the 4000 case-control cohort. Particularly, we listed the variants that are protein altering, i.e. changing the amino acid sequences of the encoded protein, changing single amino acids in the protein, or affecting the splicing of the protein-coding transcript. These variants are prioritized as they potentially change the protein structure and function. SNVs in the enhancer and promoter regions of a protein-coding gene are also detected. The annotation of these variants is more complicated as the exact location of enhancer, promoter and transcriptional factor binding sites (TFBS) and their functions are not as clear as protein-coding genes. To best annotate these variants, we incorporated information from FANTOM5 (Lizio *et al.* 2015) and ENCODE (Consortium 2012). Other variants in the exonic region of ncRNA, and highly significant variants in intronic and intergenic regions are also presented. However, the annotation of those variants is more difficult and requires further studies.

2. **Gene level**: Calculate the combined effects of a group of variants within a region, instead of each individual variants. Sequence Kernel Association Test (SKAT) (Wu *et al.* 2010) is commonly used to empower the test on rare variants. Althrough it was orginally designed for unrelated case-control study, it has been extended to familial models. We used a generalized linear mixed model F-SKAT (Yan *et al.* 2015) that is suitable for studying dichotomous traits in familial cohorts. We first tested in a traditional approach, using all rare variants within a gene region, and reported the genes that might have been most significantly affected by the rare variants. However, the genes that were tested are highly related to the number of rare variants found within its region, and the statistical significance was moderate. However, after annotation of those genes, very limited evidence of their effect on DN were found.

   Alternatively, we applied F-SKAT test on all associated SNVs that are tested as nominally significant in case-control test. That is, we first did association test on all variants within a gene region (excluding "synonymous variants" which are the variants in exonic region but do not change amino acid sequence), those with nominal significance (Odds Ration >1.5, P <0.05) were clustered together for F-SKAT test. Using this setting, we detected a group of 206 genes that are F-SKAT significant (P <0.05). Intriguingly, when we performed functional enrichment test on this group of genes, we found that the top and only significant pathway/network was the XPodNet (Warsow *et al.* 2013), which is a protein-protein interactions in the podocyte network expanded by STRING. Knowing the importance of podocyte function in the kidney, we were surprised and excited by the very specific results from an absolute unbiased approach. Even if we performed tests taking linkage disequilibrium (LD) into consideration, we observed the change of the rank of significance. However, the significance of XPodNet and the core genes within the network remained the same.

We were particular interested in several protein kinase genes *PRKCE*, *PTK2* (F-SKAT *P*=0.0037) and *PRKCI* (F-SKAT *P*=0.0085) detected by this approach. The protein kinases have been previously studied in DN (Geraldes *et al.* 2010), and a lot of evidence has shown their functions in DN or other diabetic complications (Ishii *et al.* 1998, Das Evcimen *et al.* 2007). We believe that our findings increase the credibility of the role of PRKC family in DN, and warrants more studies on these molecules in DN.

3. **Genome level**: Identify the hot-spots of variants in genome. In genome, there are regions that are more enriched in SNVs than others. We identified these regions by adapting a method that has been used in cancer research (Weinhold *et al.* 2014) and name them as recurrently mutated region (RMR). We then tested if these RMRs are over-represented in cases or controls. Among 850,137 RMRs detected in the DSPs, there were only 732 RMRs that are significantly (FDR<0.05) over-represented in cases or controls, named as DN-RMR. Most RMRs are in intergenic region. However, we observed that DN-RMRs are more commonly overlapped with functional regions of genome including 3' and 5' UTRs, enhancers and promoter regions. This suggest that associated RMRs are more likely to affect gene functions. Functional enrichment test of the genes also links the RMR overlapping genes with pathobiologically relevent pathways like ECM-receptor interactions, focal adhesions and type I diabetes.

3.1.3 Functional validation in animal models

After the comprehensive genetic tests on the DSP cohort, and possible validation on a larger unrelated case-control cohort, we did further functional tests on our top candidate genes. Zebrafish (*Danio rerio*) is a well-established animal model for functional study in kidney field. To silence genes is relatively quick and kidney phenotypes often develop fast, usually within 5-6 days post fertilization, and the transparency of embryos makes it easy to observe the gross changes in internal structures.

As a start, we tested the gene function of *abtb1* in zebrafish. Among all protein-altering variants, there was only one that was found in cases and this creates a stop codon in the transcript of the *abtb1* gene. The knockdown of *abtb1* in the Podocin-GFP zebrafish (He *et al.* 2011, Warsow *et al.* 2013) showed typical phenotypical features that associate with podocyte damage, including pericardial edema and declined expression of glomerular GFP. Furthermore, injection of human normal *abtb1* mRNA significantly rescued the edema, while mutant mRNA did not. The results suggest a functional change of mRNA due to this particular mutation.

We also generated a mouse line with the specific Arg164Ter mutation using CRISPR/Cas9 technology, and introduced diabetic traits by breeding the line with Akita mice or by streptozotocin injection. The outcome of nephropathy was not affected by the mutation in these models of diabetes. However, it is well known that many mouse models of DN

phenocopy poorly the situation in man (Azushima *et al.* 2018). Therefore, we cannot eliminate ABTB1 as a functional relevant gene to DN. Tests using other mouse models, for example that presented by (Gurley *et al.* 2018), are needed to explore further the role of the *abtb1* mutation in DN.

3.2 PAPER II: A META ANALYSIS OF TRANSCRIPTOMIC SIGNATURE OF GLOMERULAR DISEASE

The study constitutes a meta-analysis of available transcriptomic (microarray) data and proteomics data on isolated mouse, rat and human glomeruli in normal and disease states back in 2013. This was my first study in the kidney field, providing a good introduction of the back-then-current summary of studies on glomerular diseases using omics approach.

Although the project was a continuation of a previous study in the lab, I did a full literature mining of available glomerulus specific expression data, and included 3 human data sets, 12 mouse data sets, and 1 rat data set from both published and unpublished internal studies into a meta-analysis.

I established a new pipeline to analyses microarray data if raw data was available. In this way, we could minimize batch effects. However, the signal of microarray is quantified based on the image intensity of array chips. The quantification of each array is relatively independent, despite the use of control channels. Technically, it is challenging to normalize background signal among different experiments. Additionally, the datasets included in the meta-analysis were obtained by different types of microarray chips, making the baseline normalization impossible without introducing forced correlation. Therefore, we decided to normalize each data sets independently using GCRMA methods, and performed differential expression tests on each individual set of data, followed by functional enrichment tests using KEGG database (Kanehisa *et al.* 2000). The meta-analysis was then performed on the detected differentially expressed genes (DEGs) and significant pathways.

Differential expressed genes and functional relevant pathways are reported and summarized across different data sets. Although they are meaning for each individual condition, there were no direct consensus genes or pathways concluded from the meta-analysis. This suggests that different pathophysiological mechanisms are involved in different diseases and models. On the other hand, some microarray data sets have high level of background noise thus are difficult to detect biological meaning results. Also, it is not very surprising to observe no direct agreements in this type of study design. We ambitiously and ambiguously included all available transcriptomics and proteomics data without any filtration based on biological question. The hypothesis is that if there are genes or pathways that are concordantly significantly up/down regulated among multiple conditions, that gene or pathway must play a key role in the development of glomerular damages. Back then no similar study has been done in field of glomerular disease. However, with more knowledge accumulated along my study, I would adjust the experiments in following perspectives: select meta-analysis group based on their pathophysiological relevance; instead of using ORA on KEGG alone, include

more functional enrichment tests, such as GSEA (Gene Set Enrichment Analysis) (Subramanian *et al.* 2005) and other pathway databases Gene Ontology cluster (Consortium 2015), wikipathways, etc.; examine the co-expression network between different data set; instead of using a rigid threshold for significance, use a training model to include more relevant genes in each data sets to make best biological sense of the data.

With the increasing availability and decreasing price of sequencing technology, transcriptomics data produced is growing exponentially, especially for RNA-seq. The science community has made great effort for public data depository and data availability. The transcriptomics experiments are usually served for single purposes such as comparison between conditions. However, as microarray and RNA-seq data contain comprehensive panorama of the transcriptome of the particular tissue/cell/organism, mining of the available data will be a convenient and efficient way for investigation on new research questions. And meta-analysis will be one of the data mining approaches to bring new knowledge.

3.3 PAPER III: STUDY OF KIDNEY ORGANOID FROM SINGLE CELL TRANSCRIPTOMIC POINT OF VIEW

Three-dimensional (3D) organoids generated from human induced pluripotent stem cells (hiPSC) have appealing features to be used as drug discovery models. Recent studies (Freedman *et al.* 2015, Morizane *et al.* 2015, Takasato *et al.* 2015, Ciampi *et al.* 2016, Sharmin *et al.* 2016) have generated 3D kidney organoids. Based on these studies, and with advantage of CRISPR/Cas9 technology (Cong *et al.* 2013), our collaborators generated a high-throughput protocol to produce kidney organoids that carry fluorescent tags on SIX2 and NPHS1. SIX2 is a nephron progenitor marker, and NPHS1 is a mature podocyte marker. In this way, the maturation of nephron progenitors can be monitored under microscope, which facilitates research performed on this model.

Expression of podocytes, proximal tubule, endothelial and extracellular matrix markers were detected after 15-20 days of differentiation. Confocal fluorescence images of developed 3D culture showed glomerular-like structures and mature kidney markers. Moreover, electron microscopic analysis showed characteristics of mature renal podocyte.

To characterize the molecular nature of the organoid in more detail, we performed single cell transcriptomic study. The advantage of scRNA-seq is that we capture transcripts from each individual cell, so we can determine:

1) How much differentiated is an individual cell? Is it expressing progenitor markers or mature cell markers?

2) How similar is one cell to other cells? Is a group of cells developing towards the same trait of a mature cell type?

3) How similar are hiPS-derived cells to those of human adult kidney in transcriptomic level?

For question 1 and 2, we performed two independent cell-clustering methods on the scRNA-seq data. We used SC3 algorithm to cluster the cells, and visualized them using tSNE algorithm. Both algorithms consent on cluster number of 3 and the cells were assigned similarly. Using progenitor and mature cell markers, we were also able to identify the approximate cell types. There are at least two mature cell types, podocytes (expressing NPHS1) and proximal tubular cells (expressing SPP1).

Question 3 is particularly interesting since it shows the value of organoid as an experimental model to study kidney diseases. We compared the transcriptomics of podocyte and proximal tubular clusters, with the microarray data gained from micro-dissected healthy human glomerular and tubular tissue (Woroniecka *et al.* 2011). We observed high similarity between glomerular microarray data and organoid podocyte cluster, as well as tubular microarray data and organoid proximal tubular cluster, suggesting that the generated organoid model photocopies at least some features of an adult human kidney.

The limitation of single cell transcriptome approach is that due to the different nature of data, we can only approximate, but not precisely compare the similarity between human tissue and organoid. Also, to date scRNA-seq technology can only capture 10-40% of whole transcriptome (Haque *et al.* 2017), which results in significant dropout of lowly expressed genes, and inflation of randomly captured genes. The bioinformatics approach can only diminish the noise of inflation, but the dropout of certain crucial but not abundant genes cannot be rescued.

As for this study, we observed roughly 3 clusters of cells. Other cell types like endothelial cells were observed using confocal immunofluorescence microscopy, but not detected as a cell cluster in scRNA-seq. This might be due to the limited cell number sequenced, or it can be caused by the loss of sensitivity (due to noise of data) or because of selection-bias generated during the cell sorting protocol. Further experiments are needed, but my speculation is that this is likely because of the limited cell number, as endothelium cells usually have distinct transcriptomic features, and if captured, shall be identified as a separate cell type.

4 CONCLUSIONS AND FUTURE PERSPECTIVES

This thesis presents studies of kidney diseases and models, using different approaches of omics analysis, with a focus from bioinformatics perspective.

Paper I presents a genetic study using WGS on Finnish diabetic sibling cohort that is discordant for DN, and reported the novel discovery and insights into the genetics of DN. The study approach can also be an example for future genetic studies on familial materials using WGS approaches.

One straightforward approach for future study is to perform functional validation on the top candidate genes we report in the study. For example, the top genes carrying protein-altering variants or enriched for DN associated variants could be replicated in a larger cohort and by animal experiments. Another approach from bioinformatics point of view is to design studies providing information from different perspective. For example, eQTL (expression Quantitative Trait Loci) study is available for nephrotic syndrome (Gillies *et al.* 2018) but no eQTL data for DN has been reported yet. This will facilitate the annotation and validation of many variants that has been previously identified. Moreover, new studies are continuously providing new knowledge on functional annotation of gene regions, or molecular functions involved in certain pathways, etc. Mapping of the updated information into our cohort of variants could thus be one easy but effective way to gain new insights.

A summary of transcriptomics and proteomics studies in glomeruli diseases is reported in **Paper II**. Although there are no direct consensus genes or pathways concluded from the meta-analysis, the reported most affected genes and pathways may provide reference for glomerular studies. The report also shows an example for meta-analysis on publicly available omics data sets.

In **Paper III**, a high-throughput protocol of 3D kidney organoid model derived from hiPSC is generated. Utilization of scRNA-seq technology provides an unbiased approach to detect cell types, and can be a powerful tool to interpret biological meanings of the study subject.

Apart from bioinformatics challenge of high noise, the most challenging problem remains in the sample preparation step – it is a major task to capture single cells from a fresh tissue sample while keeping the minimal changes of RNA. However, I do hope and believe with the improvement of technology, investigation of molecular activities on single cell level will greatly facilitate the study in kidney field, and shed light on new understanding of the pathogenesis of kidney diseases.

5 ACKNOWLEDGEMENT

My journey of PhD studies started on July 1st 2012 when I started in Karl Tryggvason's lab as a research engineer. The journey was originally planned for only half year, but it lasted until now. The work took me from the Nordic Stockholm to tropical Singapore. Being a PhD student is definitely a unique and life-changing experience for me. The experience provides me not only with new scientific knowledge and forms the basis of my academic career, but it has also given me training as an explorer to both the inner and outer worlds – be brave to try, and be wise to correct mistakes.

The journey would never be complete without the advices, help and companion of many people:

Karl Tryggvason, I would never be here without you. Thank you for giving me the opportunity of being your graduate student. It is my great honor and pleasure to work with you. You introduced me to the scientific field, and lead me through my clumsy beginner's years. I love the discussions with you, always visionary and inspiring. Thank you for trusting me for the projects, and being patient with me when I am lost. And thank you for providing me with a great scientific environment and great group work, both in Stockholm and in Singapore.

Jaakko Patrakka, thank you for supporting and guiding me during my years in Stockholm, even allowing me and trusting me to try some experimental laboratory work. Although it was very experimental and not very productive in the end, it is a very valuable experience for me as a bioinformatician who mainly works with computers and computing. . Thank you for introducing me the basics of molecular biology of the kidney. I learnt a lot during the discussion with you and in your lab meetings.

Enrico Petretto, you have been my practical supervisor during my last two years of PhD studies I have spent in Singapore. Thank you for accepting me to your group at Duke-NUS. It is a great pleasure to work with you and your group! I love the way you work, efficient and enjoyable. Your guidance has given me a solid and "systematic" view for how a computational scientist shall be. I wish the future work with you will lead to a fruitful and pleasant journey for both of us.

Liqun He, thank you for introducing me to my current group, and for the supervision and discussion on the early projects. I wish we would have more opportunities to work together in the future! My mentor **Min Wan** in MBB**,** thank you for your genuine suggestions when I feel uncertain or upset, both in work and in life. The time I spent with you gave me a home feeling. Wish you all the best for your new start in Qingdao.

My previous supervisors for master thesis, **Daniel Dalevi** and **Marcus Bjäreland**, thank you for guiding me into the research filed. My interests for scientific research started from the project that I did with you.

Many thanks to my colleagues in Karolinska and Duke-NUS: **Anne-May Österholm**, you are a big sister to me. Thank you for providing generous care and help all the time. I really appreciate the caring messages you sent me, and dinner invitations when I feel lonely in a foreign country. You made my start of new life much easier and pleasant, both in Stockholm and in Singapore. Wish everything fantastic for your new work! **Bing He**, you are like a teacher and mentor for me. Scientific discussion with you is always informative. I learnt a lot from you and it was a great pleasure to work with you! **Patricia Rodriguez**, thank you for sharing the PhD life of the 4 years in Stockholm. Your attitude in work and life encouraged me. **Sonia Zambrano**, you are one of the bravest girls I've ever met, I love the discussion and talks with you in the lab and at our spare time. Wish all the best to you and the little one. My colleagues in Matrix, **Ann-Sofie Nilsson**, **Kan Katayama, Juha Ojala, Sergey Rodin, Sussi Virding, Olle Rengby, Ásmundur Oddsson, Sam Tryggvason** and other previous members in Matrix, thank you for sharing the knowledge and being always supportive for discussions. I also love the afterwork time with you too. My colleagues in ICMC, **Katja Möller-Hackbarth**, **Dadi Xu**, **Xiaojie Ma**, **Angelina Schwarz**, **Lwaki Ebarasi**, thank you all for sharing discussion and knowledge with me. My friends across the corridor, **Yixin Wang, Chenfei Ning, Christine Mössinger, Yi Jin, Azadeh Nilchian, Hassan Foroughi Asl**, **Aranzazu Rossignoli, Hong Li,** thank you for all the joyful chit chats during lunch and the happy hours afterwork. And of course the all time standby helps in work. My colleagues in MBB, including but not limited to: **Chad Tunell, Linda Fredriksson, Lars Jakobsson, Daniel Nyqvist, Mirela Balan, Agnieszka Martowicz, Bo Zhang, Alfredo Gimenez-Cassina, Husain Talukdar, Jongwook Hong, Guillem Genové, Sebastian Lewandowski,** and other members of Vascular Biology, thank you all for bringing together a supportive and lovely work environment.

My colleagues in the computation center in Duke-NUS, **Owen Rackham, Nathan Harmston, Sara Langley, Aida Moreno-Moral, Shiyang Liu, Huimei Chen, Uma Sangumathi Kamaraj, Amelia Tan, Sonia Chothani**, **John Ouyang, Eleni Christodoulou**, thank you for being so supportive as a group, your talent and passion have a great impact on me. It is a great pleasure to work with you. My colleagues in other group, **Miina Öhman, Yang Sun, Hwee Goon Tay, Elena Okina, Zhuhua Cai, Mien Nguyen, Kris Sigmundsson, Wei Sheng Tan, Li Yen Chong, Lynn Yap, Monica Tjin,** and colleagues in Duke-NUS, **Sebastian Schaefer, Shanshan Cheng, Alvin Ng, Mo Liu, Sujoy Ghosh, Steven Rozen, Babara Levy, Yang Wu, Chern Han Yong, Sonya Jane,** it is a great pleasure to work with you all, and thanks for the generous help you gave to me!

Many thanks for my collaborators: thanks for sharing your knowledge and experience with me, and all the help for the projec: **Niina Sandholm** and **Erkka Valo, Valma Harjutsalo, Carol Forsblom, Iiro Toppila, Mailli Parkkonen, Per-Henrik Groop** in Finland; **Qibin Li** and **Wenjuan Zhu** in Shenzhen; **Cecilia Boreström, Anna Reznichenk, Barbara Valastro, Henrik Palmgren, Anna Jonebring, Anna Svensson, Magnus Söderberg, Linda Cederblad, Jenny Nyström, Anna Collen, Ryan Hicks, Marcello Maresca, Anna Forslöw** in Göteborg.

I would like to give my special love to **Xiaonan Zhang**, **Tian Li, Xiao Tang** as my best friends in Stockholm. Thank you for always being there for me, and always supportive whenever and whatever happens. Thanks for sharing the joyful laughter and tears of pains and gains. I really enjoy and appreciate every minute spent with you. Wish you all a great happiness in the future, both in life and in career!

Also many thanks for my friends in Sweden. You all bring warmness in my heart and made best companion through the sunny summers and the cold winters. **Xiaoyuan Ren** and **Jiangrong Wang**, you are the friends that I feel I could always rely on, and I can see the happiness and fulfillment from your heart. May all your wishes come true very soon. **John Thilén**, still remembers the fast and furious trip in Iceland. Wish all the happiness for you and Tian together! **Bojing Liu, Yiqiao Wang, Xintong Jiang, Chang Liu**, I really miss all the time we spent on all the happy weekends. Thank you for being the best companion and sharing all the joyful playtime and discussion! **Wang Yue**, you are one of my oldest friends in Sweden, really glad that you got the life you want now, wish all the happiness for you! **Lidi Xu** and **Peng Zhang**, thank you for all the parties in Kungshamra, made a great memory there. My friends in Göteborg, **Qingnan Zhang** and **Zhenyu Huang**, **Hui Zheng**, **Xue Wang**, **Tiezheng Mao**, **Hao Chen**, **Huaqing Li**, **Cilan Cai, Xiaohua Liu, Peter Gu**, the first friends I met aboard, thank you for sharing our naïve and golden days together. Love the warm messages we sent from now and then. I miss you all! I am also grateful to all the friends I met in Sweden, including but **Hongya Han, Xinyan Miao, Xin Li, Jie Song, Qiang Zhang, Ming Liu, Bo Cao, Feng Wan, Geng Chen, Jingwen Wang, Jinzi Gao, Lijun Yang, Yuanjun Ma, Jiangnan Luo, Jia Mi, Chenjun Sun, Fang Wang, Jingru Wang, Yang Xuan, Qing Cheng, Xuefeng Li, Ning Xu, Zhiyun Pei,** etc. Thank you for all the precious memories.

My dear friends in Singapore, you made my journey joyful and the new place home. My dear roommate **Mi Ni Huang**, congratulations for marriage! It's such a pity that I cannot be there. Wish you great happiness with **Yang Zhou**. **Josep Relat**, my personal coach and **Marisa Ang,** my angel, I can feel the happiness when you are together, and thank you for your endless love and support. **Thye Chuan Tan, Yingying, Victor Lok, Alexandra Kong, Lee Ting Wong, Zaiquan Ong** and all the dear friends I met in ML. It was a great journey we made together, and thank you for always being there for me.

Special thanks to **Chee Yong Yee**, thank you for bringing sunshine to my life, and accompany me through my hard times. Life is much more colorful and joyful with you!

I would also like to cheer for my old friends, **Ting Zhang** and **Zhiyang Wang**. It has been 14 years since we first met, and I am so grateful that our friendship still last even though we've been living in different continents. Hope we will join together very soon.

To my family (致我的家人)：

亲爱的爸爸妈妈，谢谢你们无私的支持，让我可以在异国他乡没有后顾之忧地学习。谢谢你们无条件的爱，像风筝的线，不管我漂泊多远，也不会觉得迷失。是你们从小的教导，让我成为一个勇敢追求梦想的人。而你们的肯定与支持，一直都是我前进的动力。我会继续努力，成为一个让你们感到骄傲的人。希望你们照顾好自己，健康开心。我永远爱你们！还有亲爱的正言，谢谢你让我有动力成为一个更好的榜样，我同样珍惜与你分享的成长的经历。希望你健康快乐梦想成真！

6 REFERENCES

Adam, M., A. S. Potter and S. S. Potter (2017). **Psychrophilic proteases dramatically reduce single-cell RNA-seq artifacts: a molecular atlas of kidney development**. Development **144**(19): 3625-3632.

Adams, M. D., J. M. Kelley, J. D. Gocayne, M. Dubnick, M. H. Polymeropoulos, H. Xiao, C. R. Merril, A. Wu, B. Olde, R. F. Moreno and et al. (1991). **Complementary DNA sequencing: expressed sequence tags and human genome project**. Science **252**(5013): 1651-1656.

Afkarian, M., L. R. Zelnick, Y. N. Hall, P. J. Heagerty, K. Tuttle, N. S. Weiss and I. H. de Boer (2016). **Clinical Manifestations of Kidney Disease Among US Adults With Diabetes, 1988-2014**. JAMA **316**(6): 602-610.

Alicic, R. Z., M. T. Rooney and K. R. Tuttle (2017). **Diabetic Kidney Disease: Challenges, Progress, and Possibilities**. Clin J Am Soc Nephrol **12**(12): 2032-2045.

Alsaad, K. O. and A. M. Herzenberg (2007). **Distinguishing diabetic nephropathy from other causes of glomerulosclerosis: an update**. J Clin Pathol **60**(1): 18-26.

Anders, S., P. T. Pyl and W. Huber (2015). **HTSeq--a Python framework to work with high-throughput sequencing data**. Bioinformatics **31**(2): 166-169.

Azushima, K., S. B. Gurley and T. M. Coffman (2018). **Modelling diabetic nephropathy in mice**. Nat Rev Nephrol **14**(1): 48-56.

Baker, S. C.S. R. BauerR. P. BeyerJ. D. BrentonB. BromleyJ. BurrillH. CaustonM. P. ConleyR. ElespuruM. FeroC. FoyJ. FuscoeX. GaoD. L. GerholdP. GillesF. GoodsaidX. GuoJ. HackettR. D. HockettP. IkonomiR. A. IrizarryE. S. KawasakiT. Kaysser-KranichK. KerrG. KiserW. H. KochK. Y. LeeC. LiuZ. L. LiuA. Lucas*, et al.* (2005). **The External RNA Controls Consortium: a progress report**. Nat Methods **2**(10): 731-734.

Barbulovic-Nad, I., M. Lucente, Y. Sun, M. Zhang, A. R. Wheeler and M. Bussmann (2006). **Bio-microarray fabrication techniques--a review**. Crit Rev Biotechnol **26**(4): 237-259.

Borch-Johnsen, K., K. Norgaard, E. Hommel, E. R. Mathiesen, J. S. Jensen, T. Deckert and H. H. Parving (1992). **Is diabetic nephropathy an inherited complication?** Kidney Int **41**(4): 719-722.

Cech, T. R. and J. A. Steitz (2014). **The noncoding RNA revolution-trashing old rules to forge new ones**. Cell **157**(1): 77-94.

Ciampi, O., R. Iacone, L. Longaretti, V. Benedetti, M. Graf, M. C. Magnone, C. Patsch, C. Xinaris, G. Remuzzi, A. Benigni and S. Tomasoni (2016). **Generation of functional podocytes from human induced pluripotent stem cells**. Stem Cell Res **17**(1): 130-139.

Cirulli, E. T. and D. B. Goldstein (2010). **Uncovering the roles of rare variants in common disease through whole-genome sequencing**. Nat Rev Genet **11**(6): 415-425.

Collins, A. J.R. N. FoleyB. ChaversD. GilbertsonC. HerzogK. JohansenB. KasiskeN. KutnerJ. LiuW. St PeterH. GuoS. GustafsonB. HeubnerK. LambS. LiS. LiY. PengY. QiuT. RobertsM. SkeansJ. SnyderC. SolidB. ThompsonC. WangE. WeinhandlD. ZaunC. ArkoS. C. ChenF. DanielsJ. Ebben*, et al.* (2012). **'United States Renal Data System 2011 Annual Data Report: Atlas of chronic kidney disease & end-stage renal disease in the United States**. Am J Kidney Dis **59**(1 Suppl 1): A7, e1-420.

Cong, L., F. A. Ran, D. Cox, S. Lin, R. Barretto, N. Habib, P. D. Hsu, X. Wu, W. Jiang, L. A. Marraffini and F. Zhang (2013). **Multiplex genome engineering using CRISPR/Cas systems**. Science **339**(6121): 819-823.

Consortium, Encode Project (2012). **An integrated encyclopedia of DNA elements in the human genome**. Nature **489**(7414): 57-74.

Consortium, Gene Ontology (2015). **Gene Ontology Consortium: going forward**. Nucleic Acids Res **43**(Database issue): D1049-1056.

Crick, F. (1970). **Central dogma of molecular biology**. Nature **227**(5258): 561-563.

Crick, F. H. (1958). **On protein synthesis**. Symp Soc Exp Biol **12**: 138-163.

Das Evcimen, N. and G. L. King (2007). **The role of protein kinase C activation and the vascular complications of diabetes**. Pharmacol Res **55**(6): 498-510.

Derrien, T.R. JohnsonG. BussottiA. TanzerS. DjebaliH. TilgnerG. GuernecD. MartinA. MerkelD. G. KnowlesJ. LagardeL. VeeravalliX. RuanY. RuanT. LassmannP. CarninciJ. B. BrownL. LipovichJ. M. GonzalezM. ThomasC. A. DavisR. ShiekhattarT. R. GingerasT. J. HubbardC. NotredameJ. HarrowR. Guigo (2012). **The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression**. Genome Res **22**(9): 1775-1789.

Dobin, A., C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson and T. R. Gingeras (2013). **STAR: ultrafast universal RNA-seq aligner**. Bioinformatics **29**(1): 15-21.

Freedman, B. S.C. R. BrooksA. Q. LamH. FuR. MorizaneV. AgrawalA. F. SaadM. K. LiM. R. HughesR. V. WerffD. T. PetersJ. LuA. BacceiA. M. SiedleckiM. T. ValeriusK. MusunuruK. M. McNagnyT. I. SteinmanJ. ZhouP. H. LerouJ. V. Bonventre (2015). **Modelling kidney disease with CRISPR-mutant kidney organoids derived from human pluripotent epiblast spheroids**. Nat Commun **6**: 8715.

Geraldes, P. and G. L. King (2010). **Activation of protein kinase C isoforms and its impact on diabetic complications**. Circ Res **106**(8): 1319-1331.

Gheith, O., N. Farouk, N. Nampoory, M. A. Halim and T. Al-Otaibi (2016). **Diabetic kidney disease: world wide difference of prevalence and risk factors**. J Nephropharmacol **5**(1): 49-56.

Gillies, C. E., R. Putler, R. Menon, E. Otto, K. Yasutake, V. Nair, P. Hoover, D. Lieb, S. Li, S. Eddy, D. Fermin, M. T. McNulty, Network Nephrotic Syndrome Study, N. Hacohen, K. Kiryluk, M. Kretzler, X. Wen and M. G. Sampson (2018). **An eQTL Landscape of Kidney Tissue in Human Nephrotic Syndrome**. Am J Hum Genet **103**(2): 232-244.

Gregg, E. W., Y. Li, J. Wang, N. R. Burrows, M. K. Ali, D. Rolka, D. E. Williams and L. Geiss (2014). **Changes in diabetes-related complications in the United States, 1990-2010**. N Engl J Med **370**(16): 1514-1523.

Grun, D., A. Lyubimova, L. Kester, K. Wiebrands, O. Basak, N. Sasaki, H. Clevers and A. van Oudenaarden (2015). **Single-cell messenger RNA sequencing reveals rare intestinal cell types**. Nature **525**(7568): 251-255.

Gubitosi-Klug, R. A., R. Talahalli, Y. Du, J. L. Nadler and T. S. Kern (2008). **5-Lipoxygenase, but not 12/15-lipoxygenase, contributes to degeneration of retinal capillaries in a mouse model of diabetic retinopathy**. Diabetes **57**(5): 1387-1393.

Gurley, S. B., S. Ghosh, S. A. Johnson, K. Azushima, R. B. Sakban, S. E. George, M. Maeda, T. W. Meyer and T. M. Coffman (2018). **Inflammation and Immunity Pathways Regulate Genetic Susceptibility to Diabetic Nephropathy**. Diabetes **67**(10): 2096-2106.

Haque, A., J. Engel, S. A. Teichmann and T. Lonnberg (2017). **A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications**. Genome Med **9**(1): 75.

Harjutsalo, V., S. Katoh, C. Sarti, N. Tajima and J. Tuomilehto (2004). **Population-based assessment of familial clustering of diabetic nephropathy in type 1 diabetes**. Diabetes **53**(9): 2449-2454.

Harjutsalo, V., C. Maric, C. Forsblom, L. Thorn, J. Waden, P. H. Groop and Group FinnDiane Study (2011). **Sex-related differences in the long-term risk of microvascular complications by age at onset of type 1 diabetes**. Diabetologia **54**(8): 1992-1999.

Harrow, J.A. FrankishJ. M. GonzalezE. TapanariM. DiekhansF. KokocinskiB. L. AkenD. BarrellA. ZadissaS. SearleI. BarnesA. BignellV. BoychenkoT. HuntM. KayG. MukherjeeJ. RajanG. Despacio-ReyesG. SaundersC. StewardR. HarteM. LinC. HowaldA. TanzerT. DerrienJ. ChrastN. WaltersS. BalasubramanianB. PeiM. Tress*, et al.* (2012). **GENCODE: the reference human genome annotation for The ENCODE Project**. Genome Res **22**(9): 1760-1774.

He, B., L. Ebarasi, K. Hultenby, K. Tryggvason and C. Betsholtz (2011). **Podocin-green fluorescence protein allows visualization and functional analysis of podocytes**. Journal of the American Society of Nephrology : JASN **22**(6): 1019-1023.

He, B., A. M. Österholm, A. Hoverfält, C. Forsblom, E. E. Hjorleifsdottir, A. S. Nilsson, M. Parkkonen, J. Pitkäniemi, A. Hreidarsson, C. Sarti, A. J. McKnight, A. P. Maxwell, J. Tuomilehto, P. H. Groop and K. Tryggvason (2009). **Association of genetic variants at 3q22 with nephropathy in patients with type 1 diabetes mellitus**. Am J Hum Genet **84**(1): 5-13.

Hoen, P. A., Y. Ariyurek, H. H. Thygesen, E. Vreugdenhil, R. H. Vossen, R. X. de Menezes, J. M. Boer, G. J. van Ommen and J. T. den Dunnen (2008). **Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms**. Nucleic Acids Res **36**(21): e141.

Hon, C. C.J. A. RamilowskiJ. HarshbargerN. BertinO. J. RackhamJ. GoughE. DenisenkoS. SchmeierT. M. PoulsenJ. SeverinM. LizioH. KawajiT. KasukawaM. ItohA. M. BurroughsS. NomaS. DjebaliT. AlamY. A. MedvedevaA. C. TestaL. LipovichC. W. YipI. AbugessaisaM. MendezA. HasegawaD. TangT. LassmannP. HeutinkM. BabinaC. A. Wells*, et al.* (2017). **An atlas of human long non-coding RNAs with accurate 5' ends**. Nature **543**(7644): 199-204.

IDF, International Diabetes Federation (2015). **IDF DIABETES ATLAS, Seventh Edition 2015**. ISBN 978-2-930229-81-2.

Imperatore, G., R. L. Hanson, D. J. Pettitt, S. Kobes, P. H. Bennett and W. C. Knowler (1998). **Sib-pair linkage analysis for susceptibility genes for microvascular complications among Pima Indians with type 2 diabetes. Pima Diabetes Genes Group**. Diabetes **47**(5): 821-830.

Irizarry, R. A., B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf and T. P. Speed (2003). **Exploration, normalization, and summaries of high density oligonucleotide array probe level data**. Biostatistics **4**(2): 249-264.

Iseki, K. (2008). **Gender differences in chronic kidney disease**. Kidney Int **74**(4): 415-417.

Iseki, K., C. Iseki, Y. Ikemiya and K. Fukiyama (1996). **Risk of developing end-stage renal disease in a cohort of mass screening**. Kidney Int **49**(3): 800-805.

Ishii, H., D. Koya and G. L. King (1998). **Protein kinase C activation and its role in the development of vascular complications in diabetes mellitus**. J Mol Med (Berl) **76**(1): 21-31.

Islam, S., U. Kjallquist, A. Moliner, P. Zajac, J. B. Fan, P. Lonnerberg and S. Linnarsson (2011). **Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq**. Genome Res **21**(7): 1160-1167.

Jaitin, D. A., E. Kenigsberg, H. Keren-Shaul, N. Elefant, F. Paul, I. Zaretsky, A. Mildner, N. Cohen, S. Jung, A. Tanay and I. Amit (2014). **Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types**. Science **343**(6172): 776-779.

Jha, V., G. Garcia-Garcia, K. Iseki, Z. Li, S. Naicker, B. Plattner, R. Saran, A. Y. Wang and C. W. Yang (2013). **Chronic kidney disease: global dimension and perspectives**. Lancet **382**(9888): 260-272.

Ju, W.V. NairS. SmithL. ZhuK. SheddenP. X. K. SongL. H. MarianiF. H. EichingerC. C. BerthierA. RandolphJ. Y. LaiY. ZhouJ. J. HawkinsM. BitzerM. G. SampsonM. ThierC. SolierG. C. Duran-PachecoG. Duchateau-NguyenL. EssiouxB. SchottI. FormentiniM. C. MagnoneM. BobadillaC. D. CohenS. M. BagnascoL. BarisoniJ. LvH. ZhangH. Y. Wang*, et al.* (2015). **Tissue transcriptome-**

driven identification of epidermal growth factor as a chronic kidney disease biomarker. Sci Transl Med **7**(316): 316ra193.

Kanehisa, M. and S. Goto (2000). **KEGG: kyoto encyclopedia of genes and genomes**. Nucleic Acids Res **28**(1): 27-30.

Kim, D., G. Pertea, C. Trapnell, H. Pimentel, R. Kelley and S. L. Salzberg (2013). **TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions**. Genome Biol **14**(4): R36.

Kiselev, V. Y., K. Kirschner, M. T. Schaub, T. Andrews, A. Yiu, T. Chandra, K. N. Natarajan, W. Reik, M. Barahona, A. R. Green and M. Hemberg (2017). **SC3: consensus clustering of single-cell RNA-seq data**. Nat Methods **14**(5): 483-486.

Kivioja, T., A. Vaharautio, K. Karlsson, M. Bonke, M. Enge, S. Linnarsson and J. Taipale (2011). **Counting absolute numbers of molecules using unique molecular identifiers**. Nat Methods **9**(1): 72-74.

Kolodziejczyk, A. A., J. K. Kim, V. Svensson, J. C. Marioni and S. A. Teichmann (2015). **The technology and biology of single-cell RNA sequencing**. Mol Cell **58**(4): 610-620.

Kretzler, M. and R. Menon (2018). **Single-Cell Sequencing the Glomerulus, Unraveling the Molecular Programs of Glomerular Filtration, One Cell at a Time**. J Am Soc Nephrol **29**(8): 2036-2038.

Kukurba, K. R. and S. B. Montgomery (2015). **RNA Sequencing and Analysis**. Cold Spring Harb Protoc **2015**(11): 951-969.

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez*, et al.* (2001). **Initial sequencing and analysis of the human genome**. Nature **409**(6822): 860-921.

Liao, Y., G. K. Smyth and W. Shi (2014). **featureCounts: an efficient general purpose program for assigning sequence reads to genomic features**. Bioinformatics **30**(7): 923-930.

Lim, AKh (2014). **Diabetic nephropathy - complications and treatment**. Int J Nephrol Renovasc Dis **7**: 361-381.

Lizio, M.J. HarshbargerH. ShimojiJ. SeverinT. KasukawaS. SahinI. AbugessaisaS. FukudaF. HoriS. Ishikawa-KatoC. J. MungallE. ArnerJ. K. BaillieN. BertinH. BonoM. de HoonA. D. DiehlE. DimontT. C. FreemanK. FujiedaW. HideR. KaliyaperumalT. KatayamaT. LassmannT. F. MeehanK. NishikataH. OnoM. RehliA. SandelinE. A. Schultes*, et al.* (2015). **Gateways to the FANTOM5 promoter level mammalian expression atlas**. Genome Biol **16**: 22.

Lowe, R., N. Shirley, M. Bleackley, S. Dolan and T. Shafee (2017). **Transcriptomics technologies**. PLoS Comput Biol **13**(5): e1005457.

Luo, L., R. C. Salunga, H. Guo, A. Bittner, K. C. Joy, J. E. Galindo, H. Xiao, K. E. Rogers, J. S. Wan, M. R. Jackson and M. G. Erlander (1999). **Gene expression profiles of laser-captured adjacent neuronal subtypes**. Nat Med **5**(1): 117-122.

Ma, R. C. (2016). **Genetics of cardiovascular and renal complications in diabetes**. J Diabetes Investig **7**(2): 139-154.

Ma, R. C. and M. E. Cooper (2017). **Genetics of Diabetic Kidney Disease-From the Worst of Nightmares to the Light of Dawn?** J Am Soc Nephrol **28**(2): 389-393.

Maaten, Laurens van der and Geoffrey Hinton (2008). **Visualizing High-Dimensional Data Using t-SNE.** Journal of Machine Learning Research **2579-2605**.

Macosko, E. Z., A. Basu, R. Satija, J. Nemesh, K. Shekhar, M. Goldman, I. Tirosh, A. R. Bialas, N. Kamitaki, E. M. Martersteck, J. J. Trombetta, D. A. Weitz, J. R. Sanes, A. K. Shalek, A. Regev and S. A. McCarroll (2015). **Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets**. Cell **161**(5): 1202-1214.

Mäkinen, Ville-Petteri (2010). **Computational analysis of the metabolic phenotypes in type 1 diabetes and their associations with mortality and diabetic complications.** Thesis.

Martin, J. A. and Z. Wang (2011). **Next-generation transcriptome assembly**. Nat Rev Genet **12**(10): 671-682.

Menon, R., E. A. Otto, A. Kokoruda, J. Zhou, Z. Zhang, E. Yoon, Y. C. Chen, O. Troyanskaya, J. R. Spence, M. Kretzler and C. Cebrian (2018). **Single-cell analysis of progenitor cell dynamics and lineage specification in the human fetal kidney**. Development **145**(16).

Mishra, J., Q. Ma, A. Prada, M. Mitsnefes, K. Zahedi, J. Yang, J. Barasch and P. Devarajan (2003). **Identification of neutrophil gelatinase-associated lipocalin as a novel early urinary biomarker for ischemic renal injury**. J Am Soc Nephrol **14**(10): 2534-2543.

Moczulski, D. K., J. J. Rogus, A. Antonellis, J. H. Warram and A. S. Krolewski (1998). **Major susceptibility locus for nephropathy in type 1 diabetes on chromosome 3q: results of novel discordant sib-pair analysis**. Diabetes **47**(7): 1164-1169.

Morizane, R., A. Q. Lam, B. S. Freedman, S. Kishi, M. T. Valerius and J. V. Bonventre (2015). **Nephron organoids derived from human pluripotent stem cells model kidney development and injury**. Nat Biotechnol **33**(11): 1193-1200.

Morozova, O., M. Hirst and M. A. Marra (2009). **Applications of new sequencing technologies for transcriptome analysis**. Annu Rev Genomics Hum Genet **10**: 135-151.

Nathan, D. M., B. Zinman, P. A. Cleary, J. Y. Backlund, S. Genuth, R. Miller, T. J. Orchard, Control Diabetes, Interventions Complications Trial/Epidemiology of Diabetes and Group Complications Research (2009). **Modern-day clinical course of type 1 diabetes mellitus after 30 years' duration: the diabetes control and complications trial/epidemiology of diabetes interventions and complications and Pittsburgh epidemiology of diabetes complications experience (1983-2005)**. Arch Intern Med **169**(14): 1307-1316.

Neal, B., V. Perkovic, K. W. Mahaffey, D. de Zeeuw, G. Fulcher, N. Erondu, W. Shaw, G. Law, M. Desai, D. R. Matthews and Canvas Program Collaborative Group (2017). **Canagliflozin and Cardiovascular and Renal Events in Type 2 Diabetes**. N Engl J Med **377**(7): 644-657.

Nookaew, I., M. Papini, N. Pornputtapong, G. Scalcinati, L. Fagerberg, M. Uhlen and J. Nielsen (2012). **A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in Saccharomyces cerevisiae**. Nucleic Acids Res **40**(20): 10084-10097.

O'Shaughnessy, M. M., M. E. Montez-Rath, R. A. Lafayette and W. C. Winkelmayer (2015). **Patient characteristics and outcomes by GN subtype in ESRD**. Clin J Am Soc Nephrol **10**(7): 1170-1178.

Orchard, T. J., A. M. Secrest, R. G. Miller and T. Costacou (2010). **In the absence of renal disease, 20 year mortality risk in type 1 diabetes is comparable to that of the general population: a report from the Pittsburgh Epidemiology of Diabetes Complications Study**. Diabetologia **53**(11): 2312-2319.

Österholm, A. M., B. He, J. Pitkäniemi, L. Albinsson, T. Berg, C. Sarti, J. Tuomilehto and K. Tryggvason (2007). **Genome-wide scan for type 1 diabetic nephropathy in the Finnish population reveals suggestive linkage to a single locus on chromosome 3q**. Kidney Int **71**(2): 140-145.

Ozsolak, F. and P. M. Milos (2011). **RNA sequencing: advances, challenges and opportunities**. Nat Rev Genet **12**(2): 87-98.

Park, J., R. Shrestha, C. Qiu, A. Kondo, S. Huang, M. Werth, M. Li, J. Barasch and K. Susztak (2018). **Single-cell transcriptomics of the mouse kidney reveals potential cellular targets of kidney disease**. Science **360**(6390): 758-763.

Picelli, S., O. R. Faridani, A. K. Bjorklund, G. Winberg, S. Sagasser and R. Sandberg (2014). **Full-length RNA-seq from single cells using Smart-seq2**. Nat Protoc **9**(1): 171-181.

Pollen, A. A.T. J. NowakowskiJ. ShugaX. WangA. A. LeyratJ. H. LuiN. LiL. SzpankowskiB. FowlerP. ChenN. RamalingamG. SunM. ThuM. NorrisR. LebofskyD. ToppaniD. W. Kemp, 2ndM. WongB. ClerksonB. N. JonesS. WuL. KnutssonB. AlvaradoJ. WangL. S. WeaverA. P. MayR. C. JonesM. A. UngerA. R. KriegsteinJ. A. West (2014). **Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex**. Nat Biotechnol **32**(10): 1053-1058.

Pozhitkov, A. E., D. Tautz and P. A. Noble (2007). **Oligonucleotide microarrays: widely applied--poorly understood**. Brief Funct Genomic Proteomic **6**(2): 141-148.

Reichard, P., B. Y. Nilsson and U. Rosenqvist (1993). **The effect of long-term intensified insulin treatment on the development of microvascular complications of diabetes mellitus**. N Engl J Med **329**(5): 304-309.

Roshan, B. and R. C. Stanton (2013). **A story of microalbuminuria and diabetic nephropathy**. J Nephropathol **2**(4): 234-240.

Sandholm, N.N. Van ZuydamE. AhlqvistT. JuliusdottirH. A. DeshmukhN. W. RaynerB. Di CamilloC. ForsblomJ. FadistaD. ZiemekR. M. SalemL. T. HirakiM. PezzolesiD. TregouetE. DahlstromE. ValoN. OskolkovC. LadenvallM. L. MarcovecchioJ. CooperF. SamboA. MaloviniM. ManfriniA. J. McKnightM. LajerV. HarjutsaloD. GordinM. ParkkonenJaakko Tuomilehto FinnDiane Study GroupV. Lyssenko*, et al.* (2017). **The Genetic Landscape of Renal Complications in Type 1 Diabetes**. J Am Soc Nephrol **28**(2): 557-574.

Sanger, F. and A. R. Coulson (1975). **A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase**. J Mol Biol **94**(3): 441-448.

Satija, R., J. A. Farrell, D. Gennert, A. F. Schier and A. Regev (2015). **Spatial reconstruction of single-cell gene expression data**. Nat Biotechnol **33**(5): 495-502.

Schena, M., D. Shalon, R. W. Davis and P. O. Brown (1995). **Quantitative monitoring of gene expression patterns with a complementary DNA microarray**. Science **270**(5235): 467-470.

Seaquist, E. R., F. C. Goetz, S. Rich and J. Barbosa (1989). **Familial clustering of diabetic kidney disease. Evidence for genetic susceptibility to diabetic nephropathy**. N Engl J Med **320**(18): 1161-1165.

Sharmin, S., A. Taguchi, Y. Kaku, Y. Yoshimura, T. Ohmori, T. Sakuma, M. Mukoyama, T. Yamamoto, H. Kurihara and R. Nishinakamura (2016). **Human Induced Pluripotent Stem Cell-Derived Podocytes Mature into Vascularized Glomeruli upon Experimental Transplantation**. J Am Soc Nephrol **27**(6): 1778-1791.

Shiraki, T., S. Kondo, S. Katayama, K. Waki, T. Kasukawa, H. Kawaji, R. Kodzius, A. Watahiki, M. Nakamura, T. Arakawa, S. Fukuda, D. Sasaki, A. Podhajska, M. Harbers, J. Kawai, P. Carninci and Y. Hayashizaki (2003). **Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage**. Proc Natl Acad Sci U S A **100**(26): 15776-15781.

Silbiger, S. and J. Neugarten (2008). **Gender and human chronic renal disease**. Gend Med **5 Suppl A**: S3-S10.

Sim, G. K., F. C. Kafatos, C. W. Jones, M. D. Koehler, A. Efstratiadis and T. Maniatis (1979). **Use of a cDNA library for studies on evolution and developmental expression of the chorion multigene families**. Cell **18**(4): 1303-1316.

Su, Z., H. Fang, H. Hong, L. Shi, W. Zhang, W. Zhang, Y. Zhang, Z. Dong, L. J. Lancashire, M. Bessarabova, X. Yang, B. Ning, B. Gong, J. Meehan, J. Xu, W. Ge, R. Perkins, M. Fischer and W. Tong (2014). **An investigation of biomarkers derived from legacy microarray data for their utility in the RNA-seq era**. Genome Biol **15**(12): 523.

Subramanian, A., P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander and J. P. Mesirov (2005). **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles**. Proc Natl Acad Sci U S A **102**(43): 15545-15550.

Takasato, M., P. X. Er, H. S. Chiu, B. Maier, G. J. Baillie, C. Ferguson, R. G. Parton, E. J. Wolvetang, M. S. Roost, S. M. Chuva de Sousa Lopes and M. H. Little (2015). **Kidney organoids from human iPS cells contain multiple lineages and model human nephrogenesis**. Nature **526**(7574): 564-568.

Tang, F., C. Barbacioru, E. Nordman, B. Li, N. Xu, V. I. Bashkirov, K. Lao and M. A. Surani (2010). **RNA-Seq analysis to capture the transcriptome landscape of a single cell**. Nat Protoc **5**(3): 516-535.

Tang, F., C. Barbacioru, Y. Wang, E. Nordman, C. Lee, N. Xu, X. Wang, J. Bodeau, B. B. Tuch, A. Siddiqui, K. Lao and M. A. Surani (2009). **mRNA-Seq whole-transcriptome analysis of a single cell**. Nat Methods **6**(5): 377-382.

Trapnell, C., A. Roberts, L. Goff, G. Pertea, D. Kim, D. R. Kelley, H. Pimentel, S. L. Salzberg, J. L. Rinn and L. Pachter (2012). **Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks**. Nat Protoc **7**(3): 562-578.

Tuttle, K. R., G. L. Bakris, R. W. Bilous, J. L. Chiang, I. H. de Boer, J. Goldstein-Fuchs, I. B. Hirsch, K. Kalantar-Zadeh, A. S. Narva, S. D. Navaneethan, J. J. Neumiller, U. D. Patel, R. E. Ratner, A. T. Whaley-Connell and M. E. Molitch (2014). **Diabetic kidney disease: a report from an ADA Consensus Conference**. Diabetes Care **37**(10): 2864-2883.

Vallon, V. and S. C. Thomson (2017). **Targeting renal glucose reabsorption to treat hyperglycaemia: the pleiotropic effects of SGLT2 inhibition**. Diabetologia **60**(2): 215-225.

Velculescu, V. E., L. Zhang, B. Vogelstein and K. W. Kinzler (1995). **Serial analysis of gene expression**. Science **270**(5235): 484-487.

Venter, J. C., M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder*, et al.* (2001). **The sequence of the human genome**. Science **291**(5507): 1304-1351.

Visscher, P. M., N. R. Wray, Q. Zhang, P. Sklar, M. I. McCarthy, M. A. Brown and J. Yang (2017). **10 Years of GWAS Discovery: Biology, Function, and Translation**. Am J Hum Genet **101**(1): 5-22.

Wang, Z., M. Gerstein and M. Snyder (2009). **RNA-Seq: a revolutionary tool for transcriptomics**. Nat Rev Genet **10**(1): 57-63.

Wanner, C., S. E. Inzucchi, J. M. Lachin, D. Fitchett, M. von Eynatten, M. Mattheus, O. E. Johansen, H. J. Woerle, U. C. Broedl, B. Zinman and Empa-Reg Outcome Investigators (2016). **Empagliflozin and Progression of Kidney Disease in Type 2 Diabetes**. N Engl J Med **375**(4): 323-334.

Wapinski, O. and H. Y. Chang (2011). **Long noncoding RNAs and human disease**. Trends Cell Biol **21**(6): 354-361.

Warsow, G., N. Endlich, E. Schordan, S. Schordan, R. K. Chilukoti, G. Homuth, M. J. Moeller, G. Fuellen and K. Endlich (2013). **PodNet, a protein-protein interaction network of the podocyte**. Kidney Int **84**(1): 104-115.

Weinhold, N., A. Jacobsen, N. Schultz, C. Sander and W. Lee (2014). **Genome-wide analysis of noncoding regulatory mutations in cancer**. Nat Genet **46**(11): 1160-1165.

Wilusz, J. E., H. Sunwoo and D. L. Spector (2009). **Long noncoding RNAs: functional surprises from the RNA world**. Genes Dev **23**(13): 1494-1504.

Woroniecka, K. I., A. S. Park, D. Mohtat, D. B. Thomas, J. M. Pullman and K. Susztak (2011). **Transcriptome analysis of human diabetic kidney disease**. Diabetes **60**(9): 2354-2369.

Wu, H., A. F. Malone, E. L. Donnelly, Y. Kirita, K. Uchimura, S. M. Ramakrishnan, J. P. Gaut and B. D. Humphreys (2018). **Single-Cell Transcriptomics of a Human Kidney Allograft Biopsy Specimen Defines a Diverse Inflammatory Response**. J Am Soc Nephrol **29**(8): 2069-2080.

Wu, M. C., P. Kraft, M. P. Epstein, D. M. Taylor, S. J. Chanock, D. J. Hunter and X. Lin (2010). **Powerful SNP-set analysis for case-control genome-wide association studies**. Am J Hum Genet **86**(6): 929-942.

Wu, Zhijin, Rafael A Irizarry, Robert Gentleman, Francisco Martinez-Murillo and Forrest Spencer (2004). **A model-based background adjustment for oligonucleotide expression arrays**. Journal of the American statistical Association **99**: 909-917.

Yan, Q., H. K. Tiwari, N. Yi, G. Gao, K. Zhang, W. Y. Lin, X. Y. Lou, X. Cui and N. Liu (2015). **A Sequence Kernel Association Test for Dichotomous Traits in Family Samples under a Generalized Linear Mixed Model**. Hum Hered **79**(2): 60-68.

Young, M. D.T. J. MitchellF. A. Vieira BragaM. G. B. TranB. J. StewartJ. R. FerdinandG. CollordR. A. BottingD. M. PopescuK. W. LoudonR. Vento-TormoE. StephensonA. CaganS. J. FarndonM. Del Castillo Velasco-HerreraC. GuzzoN. RichozL. MamanovaT. AhoJ. N. ArmitageA. C. P. RiddickI. MushtaqS. FarrellD. RamplingJ. NicholsonA. FilbyJ. BurgeS. LisgoP. H. MaxwellS. Lindsay*, et al.* (2018). **Single-cell transcriptomes from human kidneys reveal the cellular identity of renal tumors**. Science **361**(6402): 594-599.

Zeisel, A., A. B. Munoz-Manchado, S. Codeluppi, P. Lonnerberg, G. La Manno, A. Jureus, S. Marques, H. Munguba, L. He, C. Betsholtz, C. Rolny, G. Castelo-Branco, J. Hjerling-Leffler and S. Linnarsson (2015). **Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq**. Science **347**(6226): 1138-1142.

Zerbino, D. R.P. AchuthanW. AkanniM. R. AmodeD. BarrellJ. BhaiK. BillisC. CumminsA. GallC. G. GironL. GilL. GordonL. HaggertyE. HaskellT. HourlierO. G. IzuoguS. H. JanacekT. JuettemannJ. K. ToM. R. LairdI. LavidasZ. LiuJ. E. LovelandT. MaurelW. McLarenB. MooreJ. MudgeD. N. MurphyV. NewmanM. Nuhn*, et al.* (2018). **Ensembl 2018**. Nucleic Acids Res **46**(D1): D754-D761.

Zhao, S., W. P. Fung-Leung, A. Bittner, K. Ngo and X. Liu (2014). **Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells**. PLoS One **9**(1): e78644.

Ziegenhain, C., B. Vieth, S. Parekh, B. Reinius, A. Guillaumet-Adkins, M. Smets, H. Leonhardt, H. Heyn, I. Hellmann and W. Enard (2017). **Comparative Analysis of Single-Cell RNA Sequencing Methods**. Mol Cell **65**(4): 631-643 e634.