Check for updates

Integrated multi-omics approaches to improve classification of chronic kidney disease

Sean Eddy[1], *Laura H. Mariani*[1] and *Matthias Kretzler*[1,2]✉

Abstract | Chronic kidney diseases (CKDs) are currently classified according to their clinical features, associated comorbidities and pattern of injury on biopsy. Even within a given classification, considerable variation exists in disease presentation, progression and response to therapy, highlighting heterogeneity in the underlying biological mechanisms. As a result, patients and clinicians experience uncertainty when considering optimal treatment approaches and risk projection. Technological advances now enable large-scale datasets, including DNA and RNA sequence data, proteomics and metabolomics data, to be captured from individuals and groups of patients along the genotype–phenotype continuum of CKD. The ability to combine these high-dimensional datasets, in which the number of variables exceeds the number of clinical outcome observations, using computational approaches such as machine learning, provides an opportunity to re-classify patients into molecularly defined subgroups that better reflect underlying disease mechanisms. Patients with CKD are uniquely poised to benefit from these integrative, multi-omics approaches since the kidney biopsy, blood and urine samples used to generate these different types of molecular data are frequently obtained during routine clinical care. The ultimate goal of developing an integrated molecular classification is to improve diagnostic classification, risk stratification and assignment of molecular, disease-specific therapies to improve the care of patients with CKD.

The diagnosis type of chronic kidney disease (CKD) and its severity is currently based on clinical features such as glomerular filtration rate (GFR), proteinuria and albuminuria, as well as histological features observed in kidney biopsy samples, if available[1–7]. However, this classification approach does not capture the diversity of molecular pathways that can lead to phenotypically similar kidney disease and therefore hampers our ability to predict long-term outcomes or to test and apply targeted therapies. Given this incomplete understanding of the molecular underpinnings of disease and the inability to predict patient responses, the dearth of positive clinical trials in nephrology is not surprising[8]. However, the past decade has been witness to a surge in the use of omics techniques[9] — in particular transcriptomics and metabolomics — to study the kidney. Such studies have led to a subsequent increase in the number of integrative multi-omics studies that have the potential to enhance our understanding of the molecular processes that initiate kidney disease, influence disease progression and mediate cell-type-specific responses to treatment. By coupling the molecular omics profiles of kidney tissues with detailed genetic and epigenetic analyses and longitudinal assessments of patient outcomes, we now have the ability to identify cell-level biomarkers and drug targets, and to assess cellular responses to external stimuli. As a result, we are now in a position to bridge information across multiple data types and data domains to gain a comprehensive molecular understanding of the kidney with which to guide development of targeted therapies for precision medicine.

This Review provides an overview of the ways in which the wealth of available data can be integrated to enhance our understanding of the mechanisms underlying kidney function and failure. We describe data integration techniques that can link diverse datasets to answer clinically relevant questions regarding the mechanisms of disease pathogenesis and progression, with examples of integrative approaches that have improved our understanding of kidney diseases from a mechanistic standpoint.

Integrating high-dimensional data

The study of kidney diseases is uniquely suited to leverage data integration techniques across several data domains, as the biosamples needed to generate comprehensive genetic, transcriptomic, proteomic and

[1]*Division of Nephrology, Department of Internal Medicine, Michigan Medicine, Ann Arbor, MI, USA.*

[2]*Department of Computational Medicine and Bioinformatics, Michigan Medicine, Ann Arbor, MI, USA.*

✉*e-mail: kretzler@med.umich.edu*

https://doi.org/10.1038/s41581-020-0286-5

Key points

- Classifying kidney diseases according to their molecular mechanisms has the potential to improve patient outcomes through the identification and development of more targeted therapeutic approaches.
- Omics data derived from kidney biopsy tissue not only have the potential to enable identification of disease mechanisms but might also enable the identification of prognostic and predictive biomarkers; this approach might be extended to enable the use of non-invasive surrogate urinary markers that reflect molecular pathways activated in the kidney.
- The success of targeting molecular pathways identified through approaches that involve the integration of various datasets has been demonstrated in clinical trials.
- Combining targeted therapies with predictive biomarkers puts nephrology in a position to test the concept of precision medicine, enabling trials to identify the right trial for the right patient at the right time.

metabolomic data[10–12] are routinely collected in clinical practice. Genetic testing, for example, is relatively common, since multiple monogenic mutations cause or modify kidney disease[13]. Integration of genetic or epigenetic data with molecular, histological and clinical datasets can help to delineate the functional relevance of genetic[14,15] and epigenetic[16,17] variations, and provide insights into how they might affect kidney function. Data obtained from kidney biopsy samples can also provide insights into disease mechanisms. This tissue can be used to generate transcriptomic data, either through bulk RNA sequencing of compartment-specific[18] or nephron segment-specific[19] tissue or through single-cell RNA sequencing (scRNA-seq)[20–22]. Tissue-based proteomic[23] and metabolomic[24] profiles can also be obtained from biopsy samples. In addition, digital images of biopsy material can be used to generate high-dimensional, descriptive and quantitative pathology data for computer-aided analysis using machine-learning approaches. Moreover, urine samples can serve as 'liquid biopsies', from which transcriptomic, proteomic and metabolomic datasets can be generated and integrated with corresponding kidney tissue omics data to identify non-invasive surrogate markers in the urine that reflect cellular processes in the kidney.

The integration of such multi-layered datasets[25] has the potential to provide insights into various aspects of kidney biology (FIG. 1). As discussed below, we can use integrative approaches to assess which networks and pathways are active in CKD to identify novel targets for therapy, whether model systems are representative of human kidney disease, and whether omics datasets generated at the time of biopsy or disease diagnosis can predict the future risk of kidney failure.

In cis
Located on the same strand of DNA as the gene or variant in question and in the case of eQTLs, located in relatively close proximity to the transcriptional start site of a gene.

Co-expression networks
Sets of genes that display similar trends in their direction of regulation across patients and are thus presumed to be co-regulated and therefore functionally related.

Uncovering CKD pathways and genes

CKD is a condition characterized by the loss of kidney function over time and can arise from a variety of aetiologies. As laboratory abnormalities and patient symptoms are often shared across aetiologies, the exact diagnosis is often unclear. Moreover, within our current aetiological categories, heterogeneity exists in the risk of progression and response to therapy, highlighting the likely underlying heterogeneity in molecular drivers of disease. Computational advances in omics data analysis approaches over the past decade have enabled the complexity of CKD to be explored through integration of various genetic and omics data. As described in the following sections, these different datasets can be aggregated into pathways or functional networks that are associated with disease phenotypes or outcomes to improve our ability to identify functionally relevant pathways[26].

The effect of genetic polymorphisms on pathway activation in kidney disease. Genomics studies have identified a multitude of genetic variants associated with kidney disease[14,15], but the functional effect of genetic variants and their potential causal role in disease is often poorly understood. One way of addressing the impact of genetic variants on disease is through identification of expression quantitative trait loci (eQTLs); eQTLs are chromosomal loci at which polymorphisms explain some of the variation in RNA or in protein expression of a particular gene product. These loci can be identified by integrating genomic sequencing data with transcriptomic or protein expression data[27–29]. Importantly, these studies have begun to provide insights into the functional consequences of genetic variations in regulating the transcriptional machinery in kidney disease.

One of the earliest studies to integrate single nucleotide polymorphism (SNP) and transcriptomic data from patients with CKD demonstrated an interplay between inflammatory responses and metabolic pathways[30]. The study considered 16 SNP loci that had been identified in a previous meta-analysis of genome-wide association study data from the CKDGen and Cohorts for Heart and Ageing Research in Genomic Epidemiology consortia as being associated with estimated GFR (eGFR). Using kidney biopsy samples from patients with different aetiologies of CKD, the researchers identified 18 candidate genes that were located on the same chromosome, and *in cis* with the SNPs, that also correlated with eGFR across CKD stages in an independent cohort — the European Renal cDNA bank. For each of the 18 genes, co-expression networks were generated using transcriptional data from individuals across and within pathological disease classifications, leading to the identification of cellular regulatory pathways that were associated with CKD. In particular, this approach revealed a signalling dichotomy between activated inflammatory and immune processes, and suppressed metabolic processes in CKD. Several of the target factors and networks identified by this process, including the transcription factor NRF2 (REFS[31–33]), and the JAK–STAT[34] and retinoid signalling pathways[35], have been clinically investigated or are currently under investigation in clinical trials across glomerular diseases. Furthermore, many of these pathways were found to be shared across diseases of varied pathological diagnoses including diabetic kidney disease (DKD), IgA nephropathy, lupus nephritis and focal segmental glomerulosclerosis (FSGS), highlighting common downstream pathways that mediate kidney injury despite different disease-initiating events across various aetiologies. Kidney-specific eQTLs have been used to identify potentially novel disease activities and targets, such as lysosomal beta A mannosidase[27], whereas compartment-specific eQTLs have helped to identify both novel targets including DAB2 and cell-specific signals[28,29].

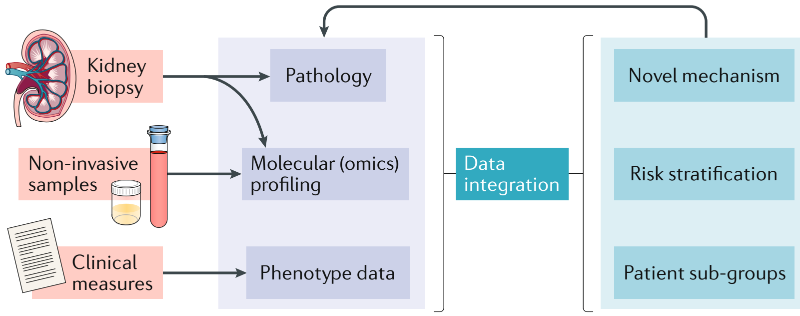

Fig. 1 | **Integration of varied data types can help to address clinically relevant questions.** Information from an individual can be captured across multiple data domains, including omics and/or pathology data from tissue, urine or blood samples as well as clinical information procured through trial or cohort enrolment. These datasets can be de-identified and coded at the individual level before they are integrated to generate associations across a cohort. Such approaches have the potential to aid risk stratification, provide insights into novel mechanisms associated with kidney disease, and aid the identification of clinically relevant subpopulations or subgroups who may benefit from targeted therapy.

The increasing availability of genetic and transcriptional datasets will further our understanding of the genetic determinants of kidney disease and help to delineate the initiating injuries and transcriptional pathways that are responsible for CKD progression in genetically susceptible individuals.

Integrating model systems with human datasets to accelerate drug development. Innovative data integration approaches have also been used to identify optimal CKD model systems with which to test potential therapeutic targets. Integration of genome-scale transcriptional data from human kidney biopsy samples with transcriptional data from preclinical CKD models has been used, for example, to define key regulatory hubs in human patients with CKD and identify corresponding model systems that are best suited for preclinical studies. For example, a 2013 study compared the differential expression of genes in the glomerulus of patients with DKD and high or low albuminuria with those of three mouse models of DKD: DBA/2 mice with streptozotocin-induced diabetes, C57BLKS db/db mice, and eNOS-deficient C57BLKS db/db mice[36]. An approximate network-level matching algorithm successfully identified gene–gene interaction networks that were common between patients and each of the three mouse models. This mapping approach has broad utility, enabling key regulatory hubs in human DKD biopsy datasets to be defined while simultaneously indicating which mouse model best mimics human disease. This study identified *STAT3* as one of the central nodes conserved across both mice and humans, implicating the JAK–STAT pathway in DKD pathogenesis. In support of this finding, expression levels of JAK–STAT pathway components were associated with disease progression and kidney failure in these patients. Further support for this finding was provided by a separate analysis that also identified upregulation of JAK and STAT pathway components in the glomerulus and tubulointerstitium of patients with DKD compared with levels in healthy living transplant donors and patients with minimal change disease[37]. A causal role for the JAK–STAT pathway in the progression of DKD was subsequently established by the demonstration that podocyte-specific overexpression of JAK2 in a mouse model of diabetes increased albuminuria, mesangial expansion, glomerulosclerosis and reduced podocyte density[38]. These phenotypes were ameliorated by a JAK1–JAK2 inhibitor. Together, these findings supported the clinical development of baricitinib, a JAK1–JAK2 inhibitor, as a treatment for DKD. Indeed, findings from a phase II study published in 2018 showed that baricitinib reduced albuminuria in patients with DKD[34]. Of interest, a separate study reported activation of JAK–STAT signalling in the peripheral blood and kidney tissue of patients with FSGS[39], potentially expanding the therapeutic opportunity for JAK–STAT intervention to other kidney diseases.

Integrating model systems with human data to identify novel disease mechanisms. The identification of molecular pathways that are specific to disease phenotypes can also be explored through data integration strategies. For example, tubulointerstitial fibrosis is a hallmark of CKD progression, irrespective of pathogenesis[40,41]. Thus, the ability to define the pathways that are responsible for the induction of this process may lead to the development of intervention strategies that are broadly applicable across disease aetiologies. One study that performed a genome-wide transcriptional analysis of tubules from biopsy samples identified differential expression of genes involved in metabolic pathways between fibrotic and normal kidneys[42]. These genes included those that encode key enzymes of fatty acid oxidation, including *CPT1A*, *CPT2*, *ACOX1*, *ACOX2* and transcriptional regulators of these genes, *PPARA* and *PPARGC1A*. Further analysis in two mouse models of kidney fibrosis showed a reduction in protein and transcript levels of fatty acid oxidation enzymes in the kidneys of fibrotic animals compared with controls, which was associated with elevated lipids and triglyceride levels in tubular epithelial cells of mice with fibrosis. In mice with folic acid-induced fibrosis, the suppression of genes and enzymes involved in fatty acid oxidation occurred prior to the onset of fibrosis, indicating that repression of these mechanisms was not simply a response to injury and was likely causal to disease. This hypothesis was confirmed by the demonstration that activation of fatty acid oxidation, through either genetic or pharmacological manipulation, protected these animals from fibrosis. Thus, not only did this approach provide novel insights into the mechanisms underlying the progression of kidney fibrosis, but it also pointed to novel interventional opportunities that may be relevant across clinical diagnostic categories.

Disruptive technologies

Novel technologies and analytical approaches have the potential to provide rapid and transformative advances in our understanding of disease processes. For example, technologies such as scRNA-seq and single nuclear RNA-seq (snRNA-seq) can provide insights into molecular processes at the cellular level, which may be of importance for studies of complex tissue, such as the kidney. Further insights into the molecular mechanisms relevant to disease and non-disease states may be gained by integrating cell-type-specific molecular data with other patient-level bulk tissue data to better understand the cellular sources of disease progression.

Kidney single-cell transcriptomics

The data integration examples outlined in the previous sections described the mapping of molecular programmes in bulk tissue samples or micro-dissected kidney compartments, in which multiple, heterogeneous cell types are combined in a single sample[43,44]. By contrast, single-cell analyses of the kidney focus on different cells that may initiate and respond differently to injury; however, analysis of cell-specific molecular pathways that are involved in disease processes is required to fully understand disease pathogenesis and response to treatment, and may allow for cell-specific drug-targeting approaches.

Efforts to resolve tissue heterogeneity have involved the isolation of single cells — for example, podocytes[43] — for molecular analysis and in silico deconvolution of genes expressed within specific cell lineages[44]. Technological advances now enable thousands of transcripts to be analysed across thousands of individual cells from complex tissues such as the kidney. Such advances provide an opportunity to redefine cellular types and subtypes within the kidney according to the transcript fingerprints of individual cells, as well as an opportunity to describe the transcriptional machinery used by these cells. Furthermore, integration of scRNA-seq and snRNA-seq datasets with anatomical and physiological data is enabling a cell-level view of differentiated organ systems, such as the kidney, to emerge. Finally, scRNA-seq profiles from kidney tissue can be used to characterize cell-type-selective expression of genes implicated in disease progression[21,28,45], including that of genes known to be monogenic causes of kidney disease[21].

The next step in the evolution of this technology requires leveraging cell-specific gene profiles to identify pathways and new molecular targets involved in disease progression. One 2020 study from the Kidney Precision Medicine Project (KPMP) (www.kpmp.org) consortium used cell-type-specific transcriptional signatures obtained by scRNA-seq from reference human nephrectomy and transplant biopsy samples to explore changes in gene expression in corresponding cell types in bulk RNA-seq datasets obtained from biopsy samples of patients with CKD from the European Renal cDNA bank and Nephrotic Syndrome Study Network (NEPTUNE) consortia. This deconvolution approach led to the identification of a gene signature that was specific to glomerular endothelial cells and was associated with exposure to immunosuppressive treatment at the time of biopsy and with long-term treatment responses in FSGS[46]. This approach highlights the potential to monitor cellular responses to treatment over time in patients, the potential to identify non-invasive biomarkers of cell-specific responses, and the potential to demonstrate the clinical relevance of cellular differences in archived patient samples with longitudinal data.

The application of scRNA-seq technologies to kidney biopsy samples is also a viable strategy to determine changes that occur in single cells with disease onset and progression. For example, an analysis of the single-cell landscape of biopsy samples from three patients with early DKD using snRNA-seq identified early alterations in cell-specific molecular pathways[47]. Consistent with a role for mesangial expansion in DKD, a number of transcriptional changes were noted in mesangial cells of patients with early DKD, including upregulation of *COL4A1*. Similarly, the use of scRNA-seq has been used to define the immunological landscape in kidneys of patients with lupus nephritis[48]. This study showed that resident immune cells within the kidney had similar gene expression characteristics to those of resident immune cells within other affected organs and identified multiple disease-specific subsets of myeloid, natural killer, and B cells and T cells. The study also showed that the gene expression profiles of myeloid and T cell clusters from urine cell pellets of patients with lupus nephritis correlated remarkably with those of immune cell clusters in the kidney, suggesting that urine could potentially be used as a non-invasive readout of myeloid and T-cell phenotypes in the kidney, and might enable, for example, assessment of the efficacy of therapeutic agents that target these cell types[49,50].

Beyond the identification and classification of cell types present within the diseased kidney, single-cell transcriptomics has potential to improve disease classification by providing insights into the cell-specific transcriptional changes that correlate with disease outcomes. In order to achieve this, however, new computational tools are needed to integrate the high-dimensional information obtained from single-cell studies (for example, gene expression data for each cell type, genes that are enriched in cell clusters, and/or ligand–receptor interactions across cell types) with clinical data to understand how cellular level molecular signals change with disease progression and remission. The KPMP–NEPTUNE-led identification of a glomerular endothelial cell phenotype linked to FSGS treatment exposure and disease progression is one example of how this approach could be implemented[46]. Such studies will lead to more complete biological models of disease progression in CKD.

Patient-specific kidney organoids

A similarly innovative strategy to study the molecular composition of the kidney and how this changes with disease has emerged from longstanding efforts of the kidney developmental research community to establish ex vivo models of nephrogenesis[51]. This work has culminated in the development of robust protocols to produce 3D kidney organoids from pluripotent stem cells, which can be used to model various aspects of organ differentiation and development[52,53]. Although much of the focus in organoid research has been on approaches to modelling developmental processes, kidney organoids also offer powerful opportunities to decipher the molecular mechanisms underlying genetic kidney disease, for example, through transcriptional profiling of single cells[54] or by assessing the response of organoids to pharmacological interventions[55] to identify disease-related transcriptional responses. The power of integrating scRNA-seq data from kidney organoids with gene expression data from bulk glomerular tissue and associated outcome data was demonstrated in a 2019 study that defined signatures of podocyte development stages during human nephrogenesis and organoid development, and moreover identified elements of these signatures that were reactivated in progressive human glomerular disease[54].

Organoids allow single-cell developmental trajectories to be overlaid in a human-based system, which is advantageous for studying kidney-disease-related genes such as *APOL1* that are absent in many common model organisms[56]. Analysis of developmental trajectories enables disease trajectories to be studied in a scalable system that is open to direct genetic or pharmacological experimental manipulation. As kidney organoids can be grown from patient-derived, inducible pluripotent stem cells (iPSCs), efforts are underway to establish drug screening and drug-discovery platforms using iPSCs from patients. These patient-derived iPSCs can provide opportunities to test therapeutic efficacy and toxicity against the patient's own genetic background[22,57–60]. In addition, transcriptional networks identified in patient-derived kidney organoids can in theory be mapped back to biopsy samples from the same patient, enabling the model system to be benchmarked.

Towards precision medicine in nephrology

The ultimate goal of obtaining a more complete understanding of the biological processes underlying CKD is to provide personalized treatment options and more accurately predict clinical outcomes for individual patients. A central tenet of precision medicine is to provide the right treatment for the right patient at the right time. This need resonates with the clinical experience of most patients with CKD, in whom disease course and response to therapy remain highly unpredictable[61,62]. Given the complexity of the kidney, we expect that the development of a prediction approach that is aligned with this precision medicine goal will require integration of diverse data types that each provide insights into biological processes and outcomes. For example, KPMP — a multi-site collaborative initiative funded by the National Institute of Diabetes and Digestive and Kidney Diseases — is aimed at leveraging these emerging technologies to define disease heterogeneity in CKD and acute kidney injury, and identify molecular pathways that can be targeted with the ultimate aim of enabling personalized care for people with acute kidney injury and CKD[63,64]. The initiative integrates tissue-level information generated by a complementary set of technologies including digital pathology-enabled pattern recognition and approaches to assess the tissue distribution pattern of mRNA, proteins and metabolites. However, despite the potential of innovative technologies such as scRNA-seq to provide biological insights, the challenges to data analysis and integration of single-cell datasets with clinical and other data are substantial. Meticulous attention to experimental processes, from tissue harvest through to data generation, data quality control, data integration and data visualization, are necessary to draw meaningful inferences[65]. In addition, advances in analytical capabilities such as machine learning are expected to push the boundaries for biologically relevant discoveries from modest sample sizes.

Urine proteomic risk predictors of CKD

Declining kidney function is typically assessed by a decline in GFR. However, although changes in GFR are sensitive to haemodynamic and glomerular changes, they do not reflect early tubular dysfunction and interstitial damage. Therefore, on its own, GFR is an incomplete indicator of CKD[66] and does not represent its underlying pathophysiology, especially during the early stages of disease. Identification of injury signatures within the urine proteome has been proposed as an alternative approach to GFR for the identification of patients with early kidney dysfunction or injury. An advantage of urine proteomic datasets is that they can be linked to transcriptomic or other omics profiles from kidney biopsy samples to identify non-invasive surrogates of signalling events in the kidney, which could be used for risk prediction[67–69], patient stratification[70] and target engagement[34,71], as well as to predict kidney function decline.

Clinical settings have the potential to benefit hugely from non-invasive biomarkers that can be used to monitor disease progression and provide patients with individualized prognoses. One study, for example, developed an integrated data-driven approach to identify biomarkers of CKD progression in urine that would be predictive of two key outcomes: decreased eGFR and kidney failure[67]. By integrating kidney biopsy transcriptome and urine protein data with baseline and longitudinal clinical follow-up data, this study identified *EGF* (which encodes the protein EGF) in the kidney as one of a number of transcripts that correlated positively with eGFR at the time of biopsy in more than 250 patients; intrarenal *EGF* was also found to correlate with urine EGF (uEGF) in two cohorts. In addition, transcripts that correlated with eGFR declined over time and were found to converge on pathways that resulted in a predicted decrease in the abundance of EGF protein in the kidney. As *EGF* mRNA was selectively expressed in the distal tubular epithelial cells and also correlated negatively with tubular atrophy and interstitial fibrosis among patients with nephrotic syndrome[72], researchers addressed whether uEGF could predict loss of kidney filtration function over time. Collectively, these studies have shown that loss of kidney *EGF* expression as determined by the uEGF-to-creatinine ratio is associated with a high risk of disease progression[72] across a wide spectrum of kidney diseases and cohorts[73–76]. These studies demonstrate the potential of integrative approaches to provide insights into disease mechanisms and outcomes.

A second example is provided by the integration of urine proteomics and clinical outcomes data to identify CKD273 as a urinary peptide classifier of CKD[69]. This classifier combines 273 urinary peptides that were identified using capillary electrophoresis coupled with mass spectrometry, and were found to be differentially expressed in the urine of patients with CKD and healthy controls. This classifier was initially developed to aid early diagnosis of CKD[77]: a score generated from the 273 peptides was inversely correlated with kidney

Weighted gene co-expression network analysis
A computational method of assessing global co-expression of genes across a dataset resulting in distinct sets of co-expression networks and modules.

Quantitative pathway score
An aggregate of individual pathway components in a single quantitative measure.

MultiPLIER
A bioinformatics approach to inferring pathway or cell-type levels leveraging bulk tissue profiles across multiple datasets.

Pathway-level information extractor
A bioinformatics method of inferring pathway or cell-type levels from a bulk tissue profile.

Latent variables
Variables that are not directly measured but are inferred from a set of observations in a dataset.

function[78] and an elevated CKD273 score was associated with progression[78,79]. Of interest, a majority of the peptide fragments in CKD273 are collagen fragments[80], which is consistent with the role of fibrosis in CKD[81]. However, although CKD273 predicts eGFR decline[69] with high sensitivity, it does so with low specificity, which may limit its use in a biomarker stratification setting.

Linking structure to molecular function

Molecular markers of specific structural changes in kidney tissue can be defined by linking specific data features, for example, expression of a gene, microRNA, protein or metabolite, to existing clinical and pathology data at the patient level. Following their identification, the relevance of these molecular associates to clinical outcomes can then be assessed. For example, an analysis of the correlation between gene expression data from kidney biopsy samples and interstitial fibrosis and tubular atrophy scores from whole-slide digital images from participants in the NEPTUNE cohort enabled the identification of genes that are associated with fibrosis, and therefore decline in kidney function[72]. Gene expression profiles that correlated with interstitial fibrosis and tubular atrophy were then grouped together to elucidate biological pathways that lead to interstitial fibrosis[72]. From these pathways, therapeutic targets and molecular biomarkers of progression may be identified.

A similar approach was used to uncover molecular mechanisms that contribute to the long-term progression of DKD. One study used compartment-specific gene expression profiles derived from kidney biopsy samples from individuals with preclinical or early DKD to determine whether transcriptional profiles at the time of biopsy were predictive of disease progression[11]. A weighted gene co-expression network analysis[82] approach was used to group genes with similar expression patterns into 11 co-expression modules, which ranged in size from 129 to 2,378 gene transcripts that were presumed to be functionally related. Co-expression modules were then independently correlated with careful morphometry of the cortical interstitial fractional volume (VvInt), albumin-to-creatinine ratio and urinary iothalamate clearance at the time of biopsy. Four co-expression modules were identified that significantly correlated with VvInt; two modules correlated positively and two correlated inversely. Pathway analysis of genes aggregated from the four co-expression modules associated with Vvlnt showed that these modules also aggregated into inflammatory and fibrosis networks (genes positively correlated with VvInt) and metabolic networks (genes negatively correlated with VvInt). Importantly, the gene co-expression modules most significantly associated with VvInt were not associated with albuminuria or GFR at the time of biopsy but were associated with DKD progression after 10 years of follow-up, including loss of GFR and increased albuminuria. Thus, this approach identified not only early predictors of disease progression but also molecular pathways that are sequentially regulated during DKD progression.

Predicting response to therapy

The above-described approaches provide an integrative view of disease models, pathways, networks and genes that are associated with development and progression of CKD. However, these approaches are not able to quantitatively compare the activation level of a specific pathway across patients or in a single patient over time — information that is necessary to predict the likelihood of response to a given therapy.

Identification of patients with a high likelihood of response to treatment could be achieved using a quantitative pathway score[83] based on genes known to be co-regulated; for example, genes regulated by common transcription factors or other regulatory elements that are targeted by an interventional drug[84,85]. Alternatively, quantitative pathway scores for transcriptional programmes within known cell types or identified using cell-type-deconvolution approaches could potentially be used to determine the contribution of specific cell types within heterogeneous bulk kidney tissue to disease outcomes in response to treatment.

As discussed earlier, quantitative scores of transcriptional signatures have been used to identify various signatures associated with kidney disease, including signatures of podocyte development that are reactivated in patients with glomerular disease[54], signatures indicative of immunometabolic pathway activation across rare kidney diseases[86], an endothelial cell signature indicative of response to steroids in patients with FSGS[46] and evidence of increased JAK–STAT signalling in patients with FSGS[39]. Of interest, the quantitative transcriptional score of JAK–STAT pathway activation in some patients with FSGS was similar to that of patients with DKD[39]. As mentioned earlier, the benefit of JAK–STAT inhibition in patients with proteinuric DKD, as demonstrated by a phase II study of baricitinib[34], suggests that opportunity exists to target this pathway in a subset of patients with FSGS, specifically those with active JAK–STAT signalling. The use of biomarkers to aid the identification of patients with JAK–STAT pathway activation and FSGS will be key to assessing drug efficacy in this disease setting.

Machine learning to predict outcomes

One challenge in the analysis of omics data is the fact that sample sizes are often limited and, consequently, clinical outcome or rates of end-point events within the cohorts are often low. These factors limit the ability to replicate findings in independent cohorts.

This limitation has been addressed with the use of a machine-learning method called MultiPLIER[87], which is an extension of the pathway-level information extractor (PLIER): an algorithm that identifies transcriptional regulatory pathways[88]. The underlying hypothesis of PLIER is that sets of co-regulated genes will be functionally related and therefore aggregate in biologically meaningful networks or gene sets. Using a machine-learning process, PLIER therefore learns gene co-expression patterns in a dataset and collapses genes into patterns identified as latent variables. Using prior knowledge (for example: Gene Ontology[89–93] or cell marker genes), some of the latent variables can be mapped to known gene sets and can themselves be used to assess statistical associations and reduce the multiple hypothesis testing penalty for large datasets. In MultiPLIER, latent variables identified through PLIER in each dataset are mapped across multiple tissues, datasets and conditions to determine whether latent variables generated across large datasets comprising tens of thousands of samples can be transferred to other, smaller datasets and to data types (for example, from RNA-seq data to microarrays). Using this approach we demonstrated that latent variables identified from large compendiums of tissues and varied data types could more robustly explain biological processes associated with disease severity when applied to rare disease datasets than models trained only on the rare disease dataset. Given the abundance of large expression data repositories that already exist, approaches such

Table 1 | **Web-based translational tools for data exploration in nephrology**

Tool	Data types	Purpose	Website
Nephroseq	Transcriptomic profiles of biopsy samples from patients with kidney disease Clinical metadata from patients including age, sex, UPCR, eGFR Transcriptomic profiles of kidneys from model systems	Identifying disease-related signatures Correlation of gene expression with clinical features	nephroseq.org
NephQTL	Compartment-specific (glomeruli and tubulointerstitium) gene expression profiles from biopsy samples of 187 participants in the NEPTUNE cohort Genotype frequency of SNPs in the NEPTUNE cohort	Compartment-specific (glomerular and tubule) eQTL discovery	nephqtl.org
Human Kidney eQTL Atlas	Compartment-specific (glomeruli and tubulointerstitium) gene expression profiles from biopsy samples of 151 participants linked to SNP data	Compartment-specific (glomerular and tubule), as well as whole kidney eQTL discovery	susztaklab.com/eqtl
Nephrocell	scRNA-seq data from kidney biopsy samples and organoids	Cell-selective gene marker identification	http://nephrocell.miktmc.org/
Kidney interactive transcriptomics	Single-cell and single nuclear RNA-seq datasets from kidney biopsies, model systems, and organoids	Cell-selective gene marker identification	humphreyslab.com/SingleCell
Mouse Kidney Single Cell Atlas	Single-cell RNA-seq datasets from kidneys of mice	Cell-selective gene marker identification	susztaklab.com/sc
HumanBase	Publicly available transcriptional data from Gene Expression Omnibus Publicly available protein interaction data Transcription factor binding preferences Ontology terms from Gene Ontology, Disease Ontology and BRENDA	Tissue-specific gene network identification Functional gene expression module identification Tissue-specific GWAS prioritization Tissue-specific gene expression predictions	https://hb.flatironinstitute.org/
Kidney Systems Biology Project	Transcriptomic, protein, Chip-seq data from model systems and renal epithelial cells	Gene- and protein-centred queries in kidney tissues, cells and segments	https://hpcwebapps.cit.nih.gov/ESBL/Database/
GUDMAP	Developmentally associated profiles from in situ hybridization, immunohistochemistry, histology and RNA sequencing	Molecular atlas of gene expression for the developing organs of the GU tract High-resolution molecular anatomy that highlights development of the GU system in mice Primary data access	gudmap.org
Rebuilding a Kidney Consortium	scRNA-seq visualizations from kidney biopsy samples, model systems and organoids Shared resources with GUDMAP Lab protocols Antibody validation Cataloguing of iPSC lines	Coordinate studies and data relevant to nephron regeneration Primary data access	rebuildingakidney.org
TransMART	Pre-packaged common bioinformatic algorithms Framework established for user-supplied data	Data sharing and standardization of bioinformatics analyses	transmartfoundation.org/

This Table contains some of the key web-based tools for translational researchers in nephrology, but is not exhaustive. eGFR, estimated glomerular filtration rate; eQTL, expression quantitative trait locus; GU, genitourinary; GWAS, genome-wide association studies; iPSC, induced pluripotent stem cells; scRNA-seq, single-cell RNA sequencing; SNP, single nucleotide polymorphism; UPCR, urine protein:creatinine ratio.

Table 2 | **Consortia-based sources of integrated datasets**

Consortium	Kidney-related research focus	Website
Accelerating Medicines Partnership	Lupus nephritis	https://www.nih.gov/research-training/accelerating-medicines-partnership-amp
BEAt DKD	Diabetic kidney disease	https://www.beat-dkd.eu/
Chronic Renal Insufficiency Cohort	Chronic kidney disease	cristudy.org
Clinical Phenotyping Resource and Biobank Core	Chronic kidney disease	https://kidneycenter.med.umich.edu/clinical-phenotyping-resource-biobank-core
CureGN	Glomerular disease	curegn.org
European Renal cDNA Bank	Chronic kidney disease	https://www.ekfs.de/en/scientific-funding/currently-funded-projects/european-renale- cdna-bank-kroner-fresenius-biopsiebank
GUDMAP	Kidney development	gudmap.org
Kidney Precision Medicine Project	Diabetic kidney disease, chronic kidney disease	kpmp.org
Nephrotic Syndrome Study Network	Focal segmental glomerulosclerosis, minimal change disease, membranous nephropathy	https://neptune-study.org/
Preventing Early Renal Loss in Diabetes	Diabetic kidney disease	http://www.perl-study.org/
ReBuilding A Kidney	Nephron repair and regeneration	rebuildingakidney.org
Transformative Research in Diabetic Kidney Disease	Diabetic kidney disease	https://www.med.upenn.edu/trident/

This Table contains some of the key consortia using mutli-scalar data integration approaches in nephrology, but is not exhaustive.

as MultiPLIER can be applied to datasets to identify key features that are most significantly associated with an outcome, even in datasets of limited sample size.

Machine-learning approaches have also been used to delineate lipidomic profiles associated with CKD outcomes[94], for example, through application of a novel machine-learning algorithm termed differential network-based enrichment analysis (DNEA) to published lipidomic profiles from two independent cohorts, the Clinical Phenotyping Resource and Biobank Core[95] and the Chronic Renal Insufficiency Cohort[96]. In that study, the researchers first performed DNEA of data from participants with early-stage (stage 2–3; $n = 79$) or late-stage (stage 3–4; $n = 135$) CKD from the Clinical Phenotyping Resource and Biobank Core. They then applied these findings to data from 200 participants in the Chronic Renal Insufficiency Cohort, 121 of whom did not experience progression of CKD and 79 of whom experienced progression to kidney failure over a mean follow-up of 6 years. This analysis enabled identification of lipid network profiles associated with advanced CKD and demonstrated that alterations in triacylglycerols and cardiolipins/phosphatidylethanolamines preceded the development of kidney failure by several years.

The advantage of the DNEA approach is the lack of a priori constructed networks, which can introduce bias into analyses of standard pathway databases owing to incomplete understanding and characterization of metabolites and their presence in pathways. The inherent disadvantage of this approach is that novel networks identified as being associated with disease require further mechanistic validation, which becomes difficult when the insights gained from such studies may not be immediately obvious. Importantly, both of the above-described machine-learning approaches involved computational integration of information across different types of sample and different data types, despite a limited number of participants, highlighting the potential of this approach to study rare diseases and other small population groups. Additional machine-learning approaches being used in kidney disease are described elsewhere[97].

Online tools

In the past decade, a number of easy-to-navigate web-based data exploration platforms have been developed that enable researchers to use a variety of bioinformatic approaches to generate molecular profiles, translate in vitro and in vivo laboratory findings, and assess the disease relevance of biological pathways, networks, potential targets and biomarkers without formal computational or bioinformatics training. Several key web-based platforms specifically for nephrology enable exploration of differential RNA expression profiles[98], eQTLs[28,29], scRNA-seq data profiles[21,46,99], gene network interactions, gene and protein co-expression, and functional analyses in a tissue-specific context[100–102] (TABLE 1). In addition, platforms that facilitate data aggregation and visualization include the GenitoUrinary Development Molecular Anatomy Project (GUDMAP)[103,104] and ReBuilding a Kidney Consortium[105]. The open-source data warehouse and analysis platform, tranSMART[106], which was originally developed in a public–private partnership, has now been adopted as a data exploration tool to enable researchers to share data in analysis-ready formats across several cohort studies. Clinical and

Translational Science Award designation for institutions by the NIH has promoted team science in translational health research and promoted collaborative studies and consortia within and across academic institutions[107], with a number of widely reported patient cohort studies generating integrated datasets (TABLE 2). The KPMP has been charged with developing data analysis, integration and outreach tools for a wide spectrum of stakeholders from patient participants to clinician scientists and drug developers.

Many of the analytical approaches described in this Review follow similar schema for presenting data online to those used by other web-based resources. Data are standardized, integrated with available meta-data, including gene or protein level information from standardized sources (for example, Entrez Gene, RefSeq, GenBank, Ensembl, UniProt and platform-specific probes to gene definitions), and then organized into an ontology. Application-specific automated pipelines that input data, standardize it and transform data to create analysis-ready data and then perform the analyses to generate data visuals and outputs.

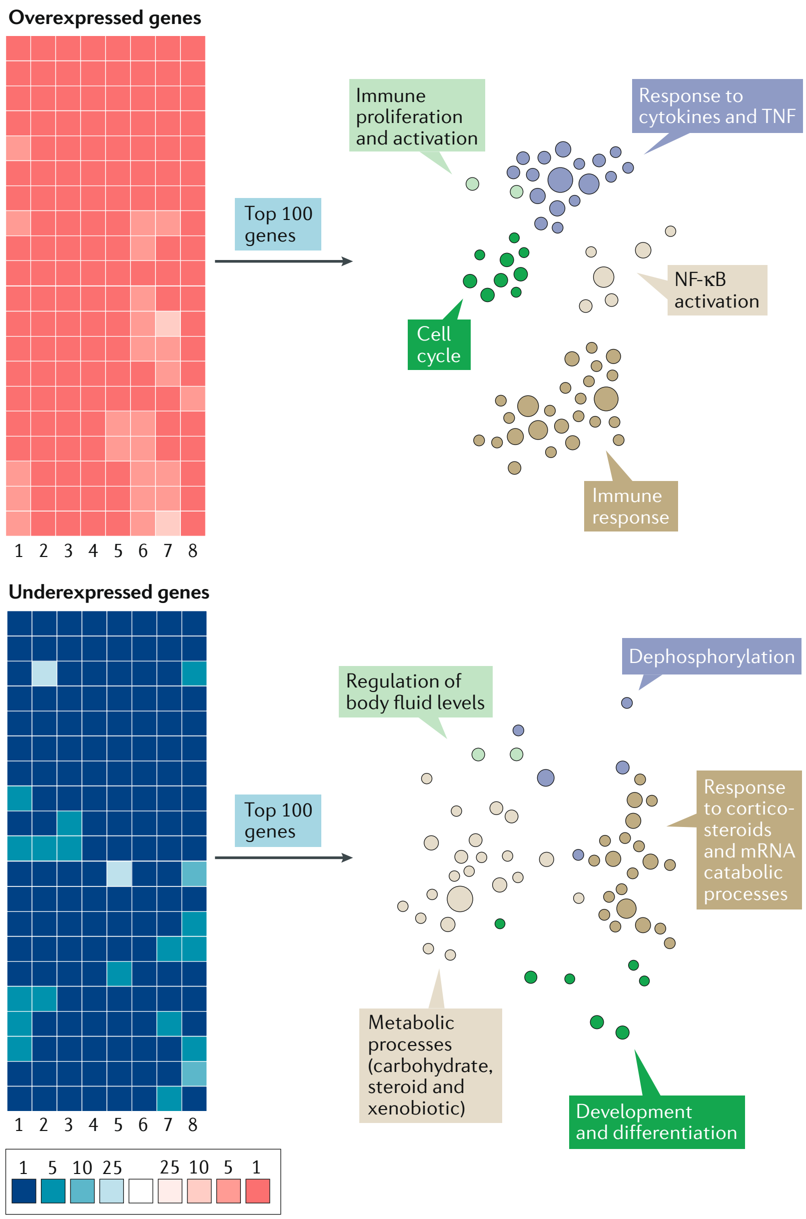

Fig. 2 | **Interrogation of public databases to provide insights into disease pathways.** Interrogation of differentially regulated genes in chronic kidney disease using Nephroseq reveals inflammatory and metabolic mechanisms associated with disease. Meta-analyses generated in Nephroseq were used to compare the gene expression profiles of glomeruli from patients with eight different types of kidney disease (1–8 corresponding to arterial hypertensive nephropathy, diabetic kidney disease, focal segmental glomerulosclerosis, IgA nephropathy, lupus nephritis, membranous nephropathy, minimal change disease and vasculitis) with the profiles of glomeruli from healthy living kidney donors. Genes that were overexpressed or underexpressed in samples from patients with kidney disease can be ranked across the eight analyses. By querying the top 100 median overexpressed and underexpressed genes in HumanBase using a kidney tissue filter, biological modules enriched in each of the gene sets can be identified. In this example, this approach reveals that the overexpressed genes are associated with molecular processes involved in inflammatory and immune networks whereas the underexpressed genes correspond to metabolic processes that are perturbed across chronic kidney disease.

The power of open data

Open data are key to the use of web-based tools for data exploration. The following is an example of how a web-based search engine, Nephroseq, can allow investigators to link their specific knowledge and interests with existing large-scale datasets to formulate and test hypotheses. Nephroseq (TABLE 1) is a web-based database that combines a number of publicly available kidney gene expression profiles with analytical tools for data mining and visualization of gene expression data. The platform is capable of creating differential expression profiles from two different analyses and performing a meta-analysis using the different differential expression profiles either within or across datasets. Thus, for example, a meta-analysis of genes identified as being differentially expressed between glomerular samples from patients with CKD of different underlying pathological conditions and healthy living donor tissue, showed that these diseases share many common molecular signals, in terms of both over- and underexpressed genes. The strikingly similar expression profile across the comparisons can be visualized using heatmaps, across all analyses (FIG. 2). To gain functional insight into the networks or mechanisms represented by these differential expression profiles, the top genes that are either overexpressed or underexpressed in the CKD samples can be analysed for tissue-specific interactions using HumanBase — a database that provides predictions about the expression, function, regulation and interactions of human gene sets. In this example, this analysis reveals that the overexpressed genes are associated with molecular processes involved in inflammatory and immune networks whereas the underexpressed genes correspond to metabolic processes that are perturbed across CKD (FIG. 2). This dichotomy of immune activation and metabolic suppression is consistent with earlier integration studies in CKD[30,42,98] and DKD, but with open data, translational researchers can now perform gene-level and network-level surveys to refine hypotheses in a matter of hours.

Although the above example illustrates the power of data sharing, privacy protection of de-identified human data is paramount, particularly in the age of genome and RNA sequencing, where re-identification of patients

Federated databases
Database management systems in which data can be securely stored while allowing users and applications to access the data through plugins.

is a real ethical and legal concern for researchers[108–110]. As demonstrated several years ago[111], de-identified individuals can be rapidly identified by interrogation of genomic and other publicly available data. Clear guidelines around data sharing and protection are one approach to addressing this issue[112]; another is the use of simulated data[113,114], which uses artificially constructed data modelled on experimental assumptions to test hypotheses. Simulated data have a clear advantage in terms of patient confidentiality. Mechanisms identified through simulated data approaches can be validated in model systems, in mechanistic studies and in targeted assays using biopsy samples from patients, while avoiding many of the potential pitfalls regarding re-identification of individuals from their genome sequence.

Finally, privacy protection might be aided by the use of federated databases, which could enable large-scale data integration studies to be performed by mounting algorithms remotely on diverse datasets, while individual datasets are kept in their protected environments. This approach involves bringing the algorithm to the data rather than bringing the data to the algorithm, and is currently being explored by several EU kidney and diabetes research networks, including Biomarker Enterprise to Attack Diabetic Kidney Disease (BEAt-DKD; www.beat-dkd.eu), and the Innovative Medicines Initiative Rhapsody (www.imi-rhapsody.eu).

Conclusions and outlook

As described in this Review, data integration approaches are key in characterizing the molecular heterogeneity in kidney disease that is not captured by pathological and clinical disease classification. As omics data become increasingly available across patient cohorts, aligning known and novel molecular definitions of disease using different technologies and platforms will be central to deciphering relevant disease- and outcome-associated biology. For instance, integration techniques that allow not only integration of two data domains (such as single-cell and bulk transcriptomic data) but also simultaneous integration of clinical data, metabolomic, histopathology and transcriptomic data, are needed to fully gain insights into disease processes. Approaches such as similarity network fusion[115] and multi-omics factor analysis[116] have been successfully used to integrate data types across domains (for example, transcriptomics, methylomics and proteomics) and platforms in other disease areas, and can help to provide a unified view of the disease. Although these approaches provide a snapshot of the disease processes, the ability to demonstrate causal effects of molecular pathways is not always possible when large datasets are assessed in isolation. In addition, caution must be taken to limit false discoveries that can be generated from integrative 'big data' approaches. Although the use of appropriate statistics certainly has a role in identifying and correcting false discoveries[26], integrating patient-derived datasets with experimental model systems can demonstrate the causality of disease-associated findings[38,117,118]. Moreover, questions relating to the causality of genes identified in genome association-wide studies can be addressed by comparing those target genes with the transcriptional profiles of cell lines following their perturbation with specific stressors, as described in the Connectivity Map (clue.io[84,85]) and in the LINCS consortium (lincsproject.org[119]) databases. Such integrative approaches are beginning to address one of the main shortcomings in this area.

As nephrology begins to move towards a precision medicine approach in which diseases are classified into molecularly defined groups, clinical trials will be able to leverage these technologies to target a more molecularly homogeneous population in whom to validate targeted therapeutics. The ultimate goal of these efforts will be to define accurate functional and molecular classification disease processes in clinical care that can improve diagnosis, prognosis and precision treatment by targeting the right patient with the right drug.

Published online 18 May 2020

1. Baigent, C. et al. Challenges in conducting clinical trials in nephrology: conclusions from a Kidney Disease-Improving Global Outcomes (KDIGO) controversies conference. *Kidney Int.* **92**, 297–305 (2017).
2. Inker, L. A. et al. KDOQI US commentary on the 2012 KDIGO clinical practice guideline for the evaluation and management of CKD. *Am. J. Kidney Dis.* **63**, 713–735 (2014).
3. Isakova, T. et al. KDOQI US commentary on the 2017 KDIGO clinical practice guideline update for the diagnosis, evaluation, prevention, and treatment of chronic kidney disease–mineral and bone disorder (CKD-MBD). *Am. J. Kidney Dis.* **70**, 737–751 (2017).
4. Lamb, E. J., Levey, A. S. & Stevens, P. E. The kidney disease improving global outcomes (KDIGO) guideline update for chronic kidney disease: evolution not revolution. *Clin. Chem.* **59**, 462–465 (2020).
5. Levey, A. S., Becker, C. & Inker, L. A. Glomerular filtration rate and albuminuria for detection and staging of acute and chronic kidney disease in adults: a systematic review. *JAMA* **313**, 837–846 (2015).
6. Lombel, R. M., Gipson, D. S. & Hodson, E. M. Kidney Disease: Improving Global Outcomes. Treatment of steroid-sensitive nephrotic syndrome: new guidelines from KDIGO. *Pediatr. Nephrol.* **28**, 415–426 (2013).
7. Lombel, R. M., Hodson, E. M. & Gipson, D. S. Kidney Disease: Improving Global Outcomes. Treatment of steroid-resistant nephrotic syndrome in children: new guidelines from KDIGO. *Pediatr. Nephrol.* **28**, 409–414 (2013).
8. Inrig, J. K. et al. The landscape of clinical trials in nephrology: a systematic review of Clinicaltrials.gov. *Am. J. Kidney Dis.* **63**, 771–780 (2014).
9. Haring, R. & Wallaschofski, H. Diving through the "-omics": the case for deep phenotyping and systems epidemiology. *OMICS* **16**, 231–234 (2012).
10. Gadegbeku, C. A. et al. Design of the nephrotic syndrome study network (NEPTUNE) to evaluate primary glomerular nephropathy by a multidisciplinary approach. *Kidney Int.* **83**, 749–756 (2013).
11. Nair, V. et al. A molecular morphometric approach to diabetic kidney disease can link structure to function and outcome. *Kidney Int.* **93**, 439–449 (2018).
12. Townsend, R. R. et al. Rationale and design of the transformative research in diabetic nephropathy (TRIDENT) study. *Kidney Int.* **97**, 10–13 (2020).
13. Groopman, E. E. et al. Diagnostic utility of exome sequencing for kidney disease. *N. Engl. J. Med.* **380**, 142–151 (2019).
14. Wuttke, M. et al. A catalog of genetic loci associated with kidney function from analyses of a million individuals. *Nat. Genet.* **51**, 957–972 (2019).
15. Hellwege, J. N. et al. Mapping eGFR loci to the renal transcriptome and phenome in the VA Million Veteran Program. *Nat. Commun.* **10**, 3842 (2019).
16. Chu, A. Y. et al. Epigenome-wide association studies identify DNA methylation associated with kidney function. *Nat. Commun.* **8**, 1286 (2017).
17. Coit, P. et al. Renal involvement in lupus is characterized by unique DNA methylation changes in naïve CD4+ T cells. *J. Autoimmun.* **61**, 29–35 (2015).
18. Cohen, C. D., Frach, K., Schlondorff, D. & Kretzler, M. Quantitative gene expression analysis in renal biopsies: a novel protocol for a high-throughput multicenter application. *Kidney Int.* **61**, 133–140 (2002).
19. Lee, J. W., Chou, C.-L. & Knepper, M. A. Deep sequencing in microdissected renal tubules identifies nephron segment-specific transcriptomes. *J. Am. Soc. Nephrol.* **26**, 2669–2677 (2015).
20. Wang, X., Park, J., Susztak, K., Zhang, N. R. & Li, M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nat. Commun.* **10**, 380 (2019).
21. Park, J. et al. Single-cell transcriptomics of the mouse kidney reveals potential cellular targets of kidney disease. *Science* **360**, 758–763 (2018).
22. Czerniecki, S. M. et al. High-throughput screening enhances kidney organoid differentiation from human pluripotent stem cells and enables automated multidimensional phenotyping. *Cell Stem Cell* **22**, 929–940 (2018).
23. Rinschen, M. M., Limbutara, K., Knepper, M. A., Payne, D. M. & Pisitkun, T. From molecules to mechanisms: functional proteomics and its application to renal tubule physiology. *Physiol. Rev.* **98**, 2571–2606 (2018).

24. Kalim, S. & Rhee, E. P. Metabolomics and kidney precision medicine. *Clin. J. Am. Soc. Nephrol.* **12**, 1726–1727 (2017).
25. Saez-Rodriguez, J., Rinschen, M. M., Floege, J. & Kramann, R. Big science and big data in nephrology. *Kidney Int.* **95**, 1326–1337 (2019).
26. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* **57**, 289–300 (1995).
27. Ko, Y. A. et al. Genetic-variation-driven gene-expression changes highlight genes with important functions for kidney disease. *Am. J. Hum. Genet.* **100**, 940–953 (2017).
28. Gillies, C. E. et al. An eQTL landscape of kidney tissue in human nephrotic syndrome. *Am. J. Hum. Genet.* **103**, 232–244 (2018).
29. Qiu, C. et al. Renal compartment–specific genetic variation analyses identify new pathways in chronic kidney disease. *Nat. Med.* **24**, 1721–1731 (2018).
30. Martini, S. et al. Integrative biology identifies shared transcriptional networks in CKD. *J. Am. Soc. Nephrol.* **25**, 2559–2572 (2014).
31. US National Library of Medicine. *ClinicalTrials.gov* https://clinicaltrials.gov/ct2/show/NCT03550443 (2019).
32. US National Library of Medicine. *ClinicalTrials.gov* https://clinicaltrials.gov/ct2/show/NCT03749447 (2019).
33. US National Library of Medicine. *ClinicalTrials.gov* https://clinicaltrials.gov/ct2/show/NCT03019185 (2019).
34. Tuttle, K. R. et al. JAK1/JAK2 inhibition by baricitinib in diabetic kidney disease: results from a phase 2 randomized controlled clinical trial. *Nephrol. Dialysis Transplant.* **33**, 1950–1959 (2018).
35. US National Library of Medicine. *ClinicalTrials.gov* https://clinicaltrials.gov/ct2/show/NCT00098020 (2017).
36. Hodgin, J. B. et al. Identification of cross-species shared transcriptional networks of diabetic nephropathy in human and mouse glomeruli. *Diabetes* **62**, 299–308 (2013).
37. Berthier, C. C. et al. Enhanced expression of Janus kinase-signal transducer and activator of transcription pathway members in human diabetic nephropathy. *Diabetes* **58**, 469–477 (2009).
38. Zhang, H. et al. Podocyte-specific JAK2 overexpression worsens diabetic kidney disease in mice. *Kidney Int.* **92**, 909–921 (2017).
39. Tao, J. L. et al. JAK-STAT signaling is activated in the kidney and peripheral blood cells of patients with focal segmental glomerulosclerosis. *Kidney Int.* **94**, 795–808 (2018).
40. Hewitson, T. D. Renal tubulointerstitial fibrosis: common but never simple. *Am. J. Physiol.-Renal Physiol.* **296**, F1239–F1244 (2009).
41. Farris, A. B. & Colvin, R. B. Renal interstitial fibrosis: mechanisms and evaluation. *Curr. Opin. Nephrol. Hypertens.* **21**, 289–300 (2012).
42. Kang, H. M. et al. Defective fatty acid oxidation in renal tubular epithelial cells has a key role in kidney fibrosis development. *Nat. Med.* **21**, 37 (2014).
43. Schroppel, B., Huber, S., Horster, M., Schlondorff, D. & Kretzler, M. Analysis of mouse glomerular podocyte mRNA by single-cell reverse transcription-polymerase chain reaction. *Kidney Int.* **53**, 119–124 (1998).
44. Ju, W. et al. Defining cell-type specificity at the transcriptional level in human disease. *Genome Res.* **23**, 1862–1873 (2013).
45. Potter, S. S. Single-cell RNA sequencing for the study of development, physiology and disease. *Nat. Rev. Nephrol.* **14**, 479–492 (2018).
46. Menon, R. et al. Single cell transcriptomics identifies focal segmental glomerulosclerosis remission endothelial biomarker. *JCI Insight* **5**, e133267 (2020).
47. Wilson, P. C. et al. The single-cell transcriptomic landscape of early human diabetic nephropathy. *Proc. Natl Acad. Sci. USA* **116**, 19619–19625 (2019).
48. Arazi, A. et al. The immune cell landscape in kidneys of patients with lupus nephritis. *Nat. Immunol.* **20**, 902–914 (2019).
49. Narain, S. & Furie, R. Update on clinical trials in systemic lupus erythematosus. *Curr. Opin. Rheumatol.* **28**, 477–487 (2016).
50. Thanou, A. & Merrill, J. T. Treatment of systemic lupus erythematosus: new therapeutic avenues and blind alleys. *Nat. Rev. Rheumatol.* **10**, 23–34 (2014).
51. Pennisi, E. Development cell by cell. *Science* **362**, 1344–1345 (2018).
52. Nishinakamura, R. Human kidney organoids: progress and remaining challenges. *Nat. Rev. Nephrol.* **15**, 613–624 (2019).
53. Little, M. H. & Combes, A. N. Kidney organoids: accurate models or fortunate accidents. *Genes. Dev.* **33**, 1319–1345 (2019).
54. Harder, J. L. et al. Organoid single cell profiling identifies a transcriptional signature of glomerular disease. *JCI Insight* **4**, pii: 122697 (2019).
55. Lemos, D. R. et al. Interleukin-1beta activates a MYC-dependent metabolic switch in kidney stromal cells necessary for progressive tubulointerstitial fibrosis. *J. Am. Soc. Nephrol.* **29**, 1690–1705 (2018).
56. Beckerman, P. & Susztak, K. APOL1: the balance imposed by infection, selection, and kidney disease. *Trends Mol. Med.* **24**, 682–695 (2018).
57. Schmidt-Ott, K. M. How to grow a kidney: patient-specific kidney organoids come of age. *Nephrol. Dial. Transpl.* **32**, 17–23 (2016).
58. Hale, L. J. et al. 3D organoid-derived human glomeruli for personalised podocyte disease modelling and drug screening. *Nat. Commun.* **9**, 5167 (2018).
59. Borestrom, C. et al. A CRISP(e)R view on kidney organoids allows generation of an induced pluripotent stem cell-derived kidney model for drug discovery. *Kidney Int.* **94**, 1099–1110 (2018).
60. Soo, J. Y. C., Jansen, J., Masereeuw, R. & Little, M. H. Advances in predictive in vitro models of drug-induced nephrotoxicity. *Nat. Rev. Nephrol.* **14**, 378–393 (2018).
61. KDIGO. KDIGO Guidelines. CKD Evaluation and Management https://kdigo.org/guidelines/ckd-evaluation-and-management/ (2012).
62. KDIGO. KDIGO Guidelines. Glomerulonephritis https://kdigo.org/guidelines/gn/ (2012).
63. Himmelfarb, J. Kidney precision medicine project: hope for the future. *ASN Kidney N.* **11**(March), 16 https://www.asn-online.org/publications/kidneynews/archives/2019/KN_2019_03_mar.pdf (2019).
64. Mariani, L. Perspectives from a junior investigator in the kidney precision medicine project. *ASN Kidney N.* **11**(March), 16–17, https://www.asn-online.org/publications/kidneynews/archives/2019/KN_2019_03_mar.pdf (2019).
65. Amezquita, R. A. et al. Orchestrating single-cell analysis with bioconductor. *Nat. Methods* **17**, 137–145 (2020).
66. Glassock, R. J. & Winearls, C. Screening for CKD with eGFR: doubts and dangers. *Clin. J. Am. Soc. Nephrol.* **3**, 1563–1568 (2008).
67. Ju, W. et al. Tissue transcriptome-driven identification of epidermal growth factor as a chronic kidney disease biomarker. *Sci. Transl. Med.* **7**, 316ra193 (2015).
68. Satirapoj, B., Pooluea, P., Nata, N. & Supasyndh, O. Urinary biomarkers of tubular injury to predict renal progression and end stage renal disease in type 2 diabetes mellitus with advanced nephropathy: a prospective cohort study. *J. Diabetes Complicat.* **33**, 675–681 (2019).
69. Pontillo, C. & Mischak, H. Urinary peptide-based classifier CKD273: towards clinical application in chronic kidney disease. *Clin. Kidney J.* **10**, 192–201 (2017).
70. Siwy, J., Klein, T., Rosler, M. & von Eynatten, M. Urinary proteomics as a tool to identify kidney responders to dipeptidyl peptidase-4 inhibition: a hypothesis-generating analysis from the MARLINA-T2D Trial. *Proteom. Clin. Appl.* **13**, 1800144 (2019).
71. US National Library of Medicine. *ClinicalTrials.gov* https://clinicaltrials.gov/ct2/show/NCT04009668 (2019).
72. Mariani, L. H. et al. Interstitial fibrosis scored on whole-slide digital imaging of kidney biopsies is a predictor of outcome in proteinuric glomerulopathies. *Nephrol. Dial. Transpl.* **33**, 310–318 (2018).
73. Wu, L. et al. Urinary epidermal growth factor predicts renal prognosis in antineutrophil cytoplasmic antibody-associated vasculitis. *Ann. Rheum. Dis.* **77**, 1339–1344 (2018).
74. Li, B. et al. Urinary epidermal growth factor as a prognostic marker for the progression of Alport syndrome in children. *Pediatr. Nephrol.* **33**, 1731–1739 (2018).
75. Azukaitis, K. et al. Low levels of urinary epidermal growth factor predict chronic kidney disease progression in children. *Kidney Int.* **96**, 214–221 (2019).
76. Yepes-Calderón, M. et al. Urinary epidermal growth factor/creatinine ratio and graft failure in renal transplant recipients: a prospective cohort study. *J. Clin. Med.* **8**, 1673 (2019).
77. Boustany, R. N., Kaye, E. & Alroy, J. Ultrastructural findings in skin from patients with Niemann-Pick disease, type C. *Pediatr. Neurol.* **6**, 177–183 (1990).
78. Argilés, À. et al. CKD273, a new proteomics classifier assessing CKD and its prognosis. *PLoS One* **8**, e62837 (2013).
79. Critselis, E. & Lambers Heerspink, H. Utility of the CKD273 peptide classifier in predicting chronic kidney disease progression. *Nephrol. Dial. Transpl.* **31**, 249–254 (2015).
80. Pontillo, C. et al. Prediction of chronic kidney disease stage 3 by CKD273, a urinary proteomic biomarker. *Kidney Int. Rep.* **2**, 1066–1075 (2017).
81. Humphreys, B. D. Mechanisms of renal fibrosis. *Annu. Rev. Physiol.* **80**, 309–326 (2018).
82. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).
83. Lee, E., Chuang, H.-Y., Kim, J.-W., Ideker, T. & Lee, D. Inferring pathway activity toward precise disease classification. *PLoS Comput. Biol.* **4**, e1000217 (2008).
84. Subramanian, A. et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell* **171**, 1437–1452 (2017).
85. Lamb, J. et al. The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* **313**, 1929–1935 (2006).
86. Grayson, P. C. et al. Metabolic pathways and immunometabolism in rare kidney diseases. *Ann. Rheum. Dis.* **77**, 1226–1233 (2018).
87. Taroni, J. N. et al. MultiPLIER: a transfer learning framework for transcriptomics reveals systemic features of rare disease. *Cell Syst.* **8**, 380–394 e384 (2019).
88. Mao, W., Zaslavsky, E., Hartmann, B. M., Sealfon, S. C. & Chikina, M. Pathway-level information extractor (PLIER) for gene expression data. *Nat. Methods* **16**, 607–610 (2019).
89. Thomas, P. D. in *The* Gene *Ontology Handbook* (eds Christophe Dessimoz & Nives Škunca) 15–24 (Springer, 2017).
90. Lewis, S. E. in *The* Gene *Ontology Handbook* (eds Christophe Dessimoz & Nives Škunca) 291–302 (Springer, 2017).
91. The Gene Ontology Consortium. Expansion of the gene ontology knowledgebase and resources. *Nucleic Acids Res.* **45**, D331–D338 (2016).
92. The Gene Ontology Consortium. The gene ontology resource: 20 years and still going strong. *Nucleic Acids Res.* **47**, D330–D338 (2018).
93. Ashburner, M. et al. Gene ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
94. Ma, J. et al. Differential network enrichment analysis reveals novel lipid pathways in chronic kidney disease. *Bioinformatics* **35**, 3441–3452 (2019).
95. Afshinnia, F. et al. Impaired beta-oxidation and altered complex lipid fatty acid partitioning with advancing CKD. *J. Am. Soc. Nephrol.* **29**, 295–306 (2018).
96. Afshinnia, F. et al. Lipidomic signature of progression of chronic kidney disease in the chronic renal insufficiency cohort. *Kidney Int. Rep.* **1**, 256–268 (2016).
97. Sealfon, R. S. G., Mariani, L. H., Kretzler, M. & Troyanskaya, O. G. Machine learning, the kidney, and genotype-phenotype analysis. *Kidney Int.* **14**, 162 (2020).
98. Martini, S., Eichinger, F., Nair, V. & Kretzler, M. Defining human diabetic nephropathy on the molecular level: integration of transcriptomic profiles with biological knowledge. *Rev. Endocr. Metab. Disord.* **9**, 267–274 (2008).
99. Wu, H. et al. Single-cell transcriptomics of a human kidney allograft biopsy specimen defines a diverse inflammatory response. *J. Am. Soc. Nephrol.* **29**, 2069–2080 (2018).
100. Greene, C. S. et al. Understanding multicellular function and disease with human tissue-specific networks. *Nat. Genet.* **47**, 569–576 (2015).
101. Krishnan, A. et al. Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. *Nat. Neurosci.* **19**, 1454–1462 (2016).
102. Zhou, J. et al. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nat. Genet.* **50**, 1171–1179 (2018).
103. McMahon, A. P. et al. GUDMAP: the genitourinary developmental molecular anatomy project. *J. Am. Soc. Nephrol.* **19**, 667–671 (2008).

104. Harding, S. D. et al. The GUDMAP database — an online resource for genitourinary research. *Development* **138**, 2845–2853 (2011).
105. Oxburgh, L. et al. (Re)building a kidney. *J. Am. Soc. Nephrol.* **28**, 1370–1378 (2017).
106. Athey, B. D., Braxenthaler, M., Haas, M. & Guo, Y. tranSMART: an open source and community-driven informatics and data sharing platform for clinical and translational research. *AMIA Jt. Summits Transl. Sci. Proc.* **2013**, 6–8 (2013).
107. Connor, E. Translating expertise: the Librarian's role in translational research. *JMLA* **106**, 137–137 (2018).
108. Dankar, F. K., Ptitsyn, A. & Dankar, S. K. The development of large-scale de-identified biomedical databases in the age of genomics-principles and challenges. *Hum. Genomics* **12**, 19 (2018).
109. Salerno, J., Knoppers, B. M., Lee, L. M., Hlaing, W. M. & Goodman, K. W. Ethics, big data and computing in epidemiology and public health. *Ann. Epidemiol.* **27**, 297–301 (2017).
110. Zarate, O. A. et al. Balancing benefits and risks of immortal data. *Hastings Cent. Rep.* **46**, 36–45 (2016).
111. Gymrek, M., McGuire, A. L., Golan, D., Halperin, E. & Erlich, Y. Identifying personal genomes by surname inference. *Science* **339**, 321–324 (2013).
112. Chico, V. The impact of the general data protection regulation on health research. *Br. Med. Bull.* **128**, 109–118 (2018).
113. Sarkar, H., Srivastava, A. & Patro, R. Minnow: a principled framework for rapid simulation of dscRNA-seq data at the read level. *Bioinformatics* **35**, i136–i144 (2019).
114. Chung, R.-H. & Kang, C.-Y. A multi-omics data simulator for complex disease studies and its application to evaluate multi-omics data analysis methods for disease classification. *GigaScience* **8**, giz045 (2019).
115. Wang, B. et al. Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods* **11**, 333–337 (2014).
116. Argelaguet, R. et al. Multi-omics factor analysis — a framework for unsupervised integration of multi-omics data sets. *Mol. Syst. Biol.* **14**, e8124 (2018).
117. Pedigo, C. E. et al. Local TNF causes NFATc1-dependent cholesterol-mediated podocyte injury. *J. Clin. Invest.* **126**, 3336–3350 (2016).
118. Mitrofanova, A. et al. Hydroxypropyl-beta-cyclodextrin protects from kidney disease in experimental Alport syndrome and focal segmental glomerulosclerosis. *Kidney Int.* **94**, 1151–1159 (2018).
119. Keenan, A. B. et al. The library of integrated network-based cellular signatures NIH program: system-level cataloging of human cells response to perturbations. *Cell Syst.* **6**, 13–24 (2018).

Acknowledgements

Support for the work described in this Review is provided by the George M. O'Brien Michigan Kidney Translational Core Center, funded by NIH/National Institute of Diabetes and Digestive and Kidney Diseases grant 2P30-DK-081943. Nephroseq is supported as part of the applied systems biology core by the University of Michigan George M. O'Brien Michigan Kidney Translational Core Center.

Author contributions

All authors contributed to discussing the article's content, writing and reviewing/editing of the manuscript before submission.

Competing interests

M.K. holds a US patent, PCT/EP2014/073413 "Biomarkers and methods for progression prediction for chronic kidney disease". S.E. and L.H.M. declare no competing interests.

Publisher's note

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

RELATED LINKS

Attack Diabetic Kidney Disease: www.beat-dkd.eu
Connectivity Map: clue.io
Lincs Consortium: lincsproject.org
Rhapsody: www.imi-rhapsody.eu
The Kidney Precision Medicine Project: www.kpmp.org