A functional landscape of chronic kidney disease entities from public transcriptomic data

Ferenc Tajti[1], Asier Antoranz[2,3], Mahmoud M. Ibrahim[1,4], Hyojin Kim[1], Francesco Ceccarelli[1], Christoph Kuppe[4], Leonidas G. Alexopoulos[2,3], Rafael Kramann[4] & Julio Saez-Rodriguez[1,5,*]

[1] RWTH Aachen University, Faculty of Medicine, Joint Research Centre for Computational Biomedicine (JRC-COMBINE), Aachen, Germany

[2] National Technical University of Athens, Greece

[3] ProtATonce Ltd, Athens, Greece

[4] Division of Nephrology and Clinical Immunology, Faculty of Medicine, RWTH Aachen University, Aachen, Germany

[5] European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Cambridge CB10 1SD, UK

* Corresponding author: saezrodriguez@gmail.com

Abstract

To develop efficient therapies and identify novel early biomarkers for chronic kidney disease (CKD) an understanding of the molecular mechanisms orchestrating it is essential. We here set out to understand how differences in CKD origin are reflected in gene regulatory mechanisms. To this end, we collected and integrated publicly available human-kidney glomerular microarray gene expression data for nine kidney disease entities that account for a majority of CKD worldwide [Focal segmental glomerulosclerosis (FSGS), Minimal Change Disease (MCD), FSGS-MCD, IgA nephropathy (IgAN), Lupus nephritis (LN), Membranous glomerulonephropathy (MGN), Diabetic nephropathy (DN), Hypertensive nephropathy (HN) and Rapidly progressive glomerulonephritis (RPGN)]. We included data from five distinct studies and compared glomerular gene expression profiles to that of non-tumor part of kidney cancer nephrectomy tissues. A major challenge was the integration of the data from different sources, platforms and conditions, that we mitigated with a bespoke stringent procedure. This allowed us to perform a global transcriptome-based delineation of different kidney disease entities, obtaining a landscape of their similarities and differences based on the genes that acquire a consistent differential expression between each kidney disease entity and tumor nephrectomy. Furthermore, we derived functional insights by inferring signaling pathway and transcription factor activity from the collected gene expression data, and identified potential drug candidates based on expression signature matching. These results provide a foundation to comprehend the specific molecular mechanisms underlying different kidney disease entities, that can pave the way to identify biomarkers and potential therapeutic targets.

1. Introduction

Chronic Kidney Disease (CKD) is a major public health burden affecting around 10 % of the population in the western world. There is no specific therapy to slow down kidney functional decline and prevent progression to end-stage renal disease (ESRD) for the vast majority of kidney diseases. Thus patients face dialysis or kidney transplantation. However, mortality and morbidity on dialysis is high and transplant wait times number in years. Furthermore, dialysis patients consume dramatic proportions of healthcare budgets [1]. Despite of this, CKD/ESRD is not receiving as much attention as other diseases, and research-funding is much lower (e.g. 100 fold less per patient than HIV [2]).

The origin of CKD is heterogenous and has slowly changed in recent years due to an aging population with increased number of patients with hypertension and diabetes. Major contributors to worldwide CKD that are studied here are Diabetic nephropathy (DN) and Hypertensive nephropathy (HN). Other contributors are immune diseases such as Lupus Nephritis (LN) and glomerulonephritides including IgA nephropathy (IgAN), Membranous glomerulonephropathy (MGN), Minimal Change Disease (MCD) as wells as Focal Segmental Glomerulosclerosis (FSGS) and Rapidly progressive glomerulonephritis (RPGN). While various other diseases such as hereditary diseases as autosomal dominant polycystic kidney disease, chronic infections, and toxins, among others, are also contributing to the prevalence of CKD the data presented here is focusing on the above mentioned diseases.

Regardless of the type of initial injury to the kidney the stereotypic response to chronic repetitive injury is scar formation and fibrosis with subsequent kidney functional decline. Scar forms in the tubulo-interstitium as tubulo-interstitial fibrosis and in the glomerulus referred to as glomerulosclerosis. Despite this stereotypic response that involves inflammation and expansion of scar secreting myofibroblasts the initiating stimuli are quite heterogeneous, ranging from an auto-immunological process in LN to poorly controlled blood glucose levels in DN. A better understanding of similarities and differences in the complex molecular process orchestrating disease initiation and progression will guide the development of novel targeted therapeutics.

A powerful tool to understand and model the molecular basis of diseases is the analysis of genome-wide gene expression data. This has been applied in the context of various kidney diseases contributing to CKD [3–7], and most studies are available in the resource NephroSeq. However, to the best of our knowledge, no study so far has combined these data sets to build a comprehensive landscape of the molecular alterations underlying different kidney diseases contributing to a majority of CKD prevalence. In this study, we set out to build such a data set. We collected data from five large studies with gene expression data from kidney biopsies of patients of eight different glomerular disease entities leading to CKD (from now on CKD entities), FSGS, MCD, IgAN, LN, MGN, DN, HN and RPGN. We normalized the data with bespoke stringent procedures, which allowed us to study the similarities and differences among these entities in terms of deregulated genes, pathways, and transcription factors, as well as to identify drugs that revert their expression signatures and thereby might be useful to treat them.

2. Results

2.1. Assemble of a pan-CKD collection of gene expression profiles of patients

We searched in Nephroseq (www.nephroseq.org) and Gene Expression Omnibus (GEO) [8,9] and identified five available studies - GSE20602 [10]; GSE32591 [11]; GSE37460 [11]; GSE47183 [12,13]; GSE50469 [14] (see section 4.1.) - with human microarray gene expression data for nine different glomerular disease entities: FSGS, MCD, IgAN, LN, MGN, DN, HN and RPGN, as well as healthy tissue and non-tumor part of kidney cancer nephrectomy tissues as controls (Figure 1A and B). In addition, for one dataset, patients were labeled as an overlap of FSGS and MCD (FSGS-MCD) and we left it as such. These studies were generated in two different microarray platforms. To be able to jointly analyze and compare the different CKD entities, we performed a stringent preprocessing and normalization procedure that involved quality control assessment, either cyclic loess normalization or YuGene transformation and a batch effect mitigation procedure (see Methods and Supplementary material), thus at the end we kept 6289 genes from 199 samples in total. From the two potential controls, healthy tissue and cancer nephrectomy, we chose the latter as reference for further analysis as the batch mitigation removed a large number of genes for the healthy tissue.

2.2. Technical heterogeneity across samples

We first examined the similarities among the samples to see the extent of potential batch effects. Data does not primarily cluster by study source or platform, which can be attributed to our batch mitigation procedure (Figure 1C, Sup. Figure 1), but the technical heterogeneity of the samples is still present, most probably due to the batch-group imbalance (see section 4.4. and Sup. Figure 1). Samples from RPGN and FSGS-MCD conditions seem to be more affected by platform-specific batch effects than samples from other conditions, most probably due to the unbalanced distribution of samples: RPGN and FSGS-MCD samples are exclusively represented in platform Affymetrix Human Genome U133 Plus 2.0 Array (GPL570). In addition both of these are the subjects of a singular study, so that the batch effect mitigation procedure was not feasible to be conducted on them. Thus, the relatively lower Spearman's correlation coefficients - 0.6 vs. for the rest - could be attributed to this uneven sample distribution in microarray platforms.

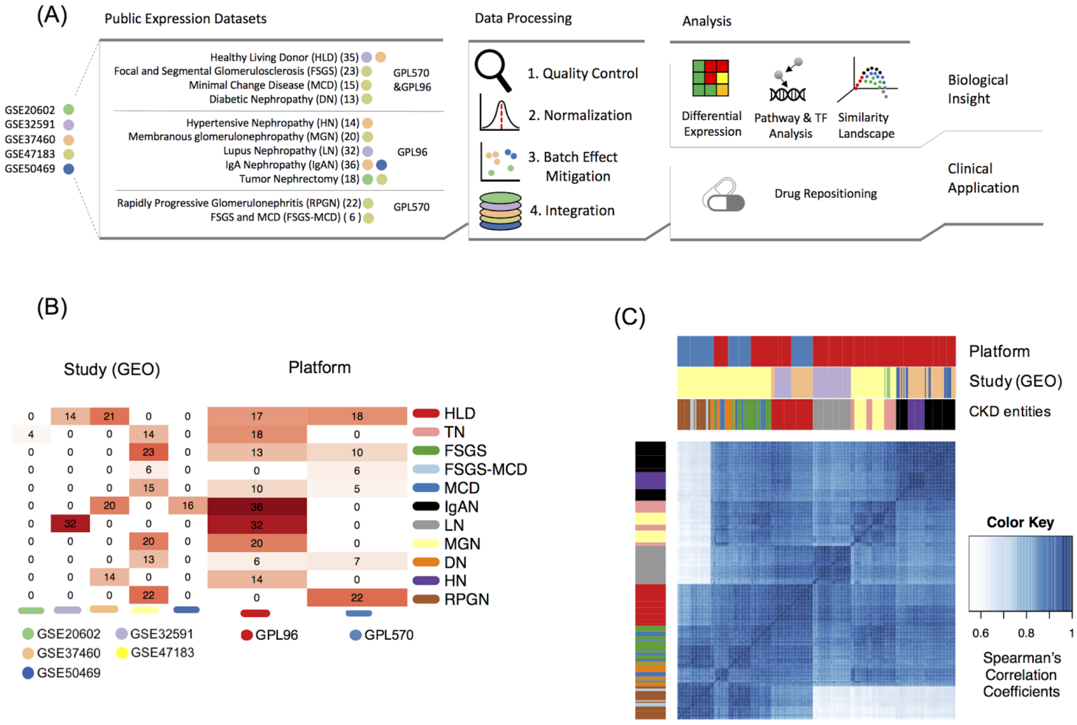


Figure 1. *(A) Flow of analysis followed in this study. (B) Heatmap of the distribution of samples across studies and microarray platforms. (C) Hierarchical clustering of the arrays based on gene expression Spearman's correlation coefficients.*

2.3. Biological heterogeneity of CKD entities

Keeping in mind the potential biases introduced by the technical heterogeneity, we set out to find molecular differences among glomerular CKD entities. For this purpose, we analyzed only the expression data obtained for glomeruli. First, we calculated the differential expression of individual genes between the different CKD entities and TN fitting linear models for microarray data [15,16]. From the 6289 genes included in the integrated dataset, 1791 showed significant differential expression (|logFC| > 1, p-value < 0.05) in at least one CKD entity. RPGN was the

CKD entity with the largest number of significantly differentially expressed genes (885), while MCD was the one with least (75). 12 genes showed significant differential expression across all the CKD entities (AGMAT, ALB, BHMT2, CALB1, CYP4A11, FOS, HAO2, HMGCS2, MT1F, MT1G, PCK1, SLC6A8). Interestingly, all these genes were underexpressed across all the CKD entities (none of them overexpressed for any CKD entity). In contrast, QKI and LYZ genes were significantly overexpressed in HN, IgAN, and LN, while significantly underexpressed in FSGS-MCD, and RPGN (and DN for QKI). 107 different genes were significantly differentially expressed in at least 6 CKD entities (Figure 2A).

To better comprehend the divergence and the similitude of the CKD samples we applied diffusion maps, as an attempt to reveal the underpinning geometric structure of the glomerular CKD transcriptomics data (Figure 2B). Specifically, we asked how the distinct CKD entities localised with respect to each other based on a common set of differentially expressed genes with regard to the expression profile of tumor nephrectomy. Thus, the diffusion distance of a given CKD entity with regard to tumor nephrectomy implies the extent of expression profile divergence from that control.

The most distant condition from tumor nephrectomy is RPGN, which is arguably the most drastic kidney disease condition with the most rapid kidney functional decline among the included kidney disease entities. Interestingly, healthy donor samples are distinct from tumor nephrectomy samples despite the fact that the tissue fragments resected from the patients with cancer were non-cancerous. This might be explained by either minor contamination with cancer cells or alternatively an effect of the tumor itself on the non-cancerous kidney tissue such as e.g. immune cell infiltration. DN and LN are in close proximity to RPGN, whereas HN is localised near IgAN. Differences are harder to asses in the middle of the diffusion map, but were visible when plotting the dimension components pair-wise (Sup. Figure 2). For instance, MCD samples spans from a point proximal to tumor nephrectomy and ends near FSGS, but some of its samples are in close proximity to MGN or even hypertensive nephropathy. While it makes sense that MCD as a relatively mild disease, without any morphological changes that can be detected by light microscopy, is relatively close to the control groups of TN and HLD, it remains unclear why other disease entities spread widely in the diffusion map. Unfortunately, the data we used does not include information about disease severity which might help to explain this heterogeneity with early stage disease closer to the control groups and late stage disease closer to RPGN. Dimension component 1 (DC1) seems to offer a focus on the dissimilarity between the two reference conditions, tumor nephrectomy and healthy living donor from the CKD entities. Dimension component 2 (DC2) provides more insight into the disparity of the reference conditions. Dimension component 3 (DC3) discerns the subtle geometrical manifestation of the distinct CKD entities with regard to each other. In summary, using diffusion maps we find clear differences in the global expression profiles of the CKD entities.

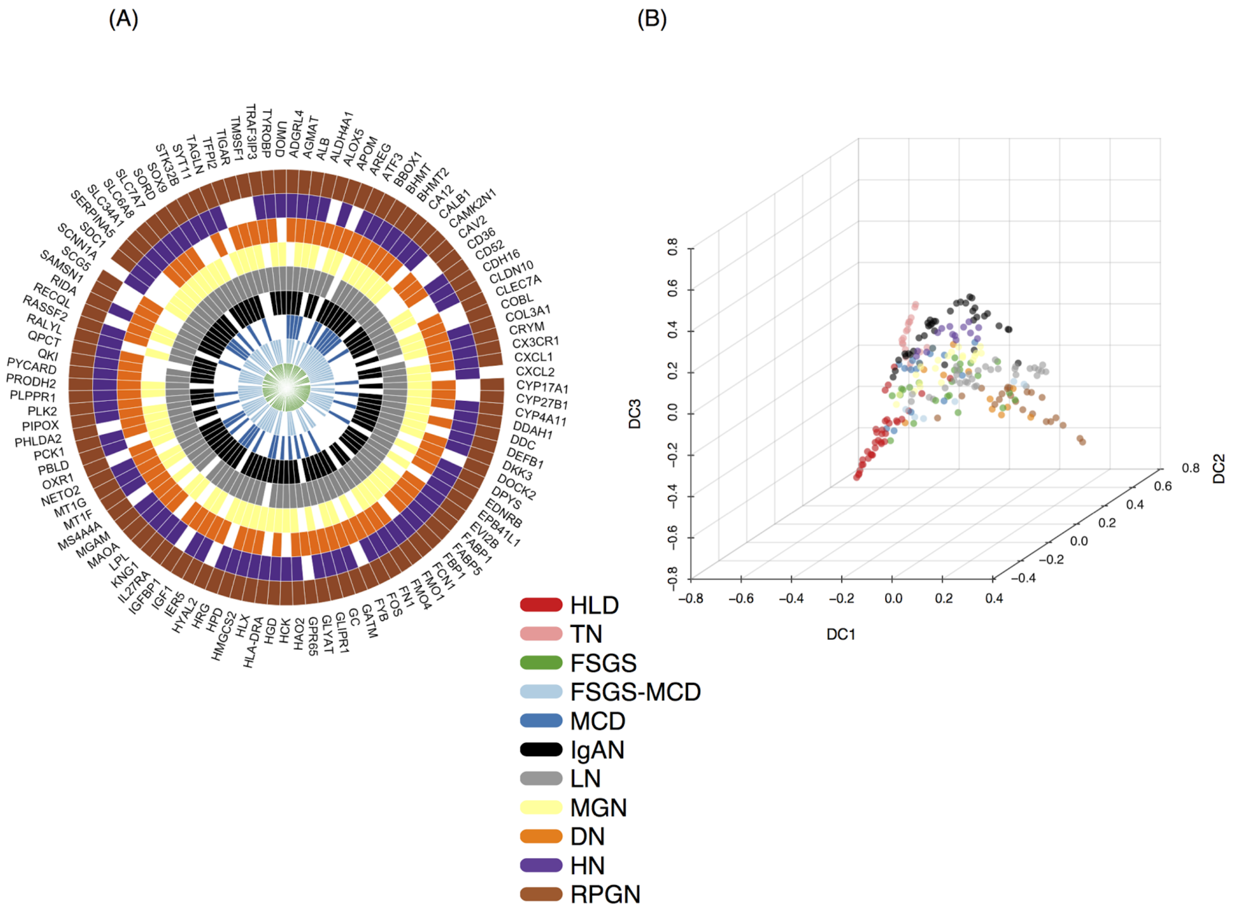


Figure 2. *(A) Radial heatmap of consistently differentially expressed genes across six or more disease entities (up- or down-regulation). (B) Diffusion map of CKD entities reveals the underpinning geometric structure of the glomerular CKD transcriptomics data.*

2.4. Transcription factor activity in CKD entities

To further characterize the differences among the CKD entities, we performed various functional analyses. First, we assessed the activity of transcription factors (TFs; Figure 3), based the levels of expression of their known targets (see Methods). Changes in the target genes provides superior estimates of the TF activity than the expression level of the transcription factor itself [17,18] (Figure 3). We found 10 TFs differentially regulated in at least one CKD entity (Figure 3). Furthermore, we correlated the identified TF's activities with the expression of those genes, that are encoding for these TFs. The idea is that, while factors with negative correlation are potentially acting as repressors, whereas those with positive correlation are acting as activators. Those with no correlation indicate factors whose activity is significantly modulated using post-translational modifications or factors whose regulons or expression measurements are unconfident. For instance, Interferon regulatory factor-1 (IRF1) is significantly enriched in Lupus nephritis and moderately correlated (Spearman's rank-based correlation coefficient of 0.624) with the expression level of the gene encoding for IRF1. This suggests an as of yet undiscovered potential role of IRF1 as a transcriptional activator in Lupus nephritis. In addition, IRF1's transcriptional activity was elevated in LN with respect to the rest of the conditions. The activity of the upstream stimulatory factor 2 (USF2) - a basic helix-loop-helix (bHLH) TF [19] - is estimated to be significantly depleted in MCD compared to the rest of the conditions. Interestingly, USF2's estimated activity across the CKD entities is inversely correlated - Spearman's rho (r_s = -0.867) - with the expression level of the gene USF2, that is encoding for the TF USF2. Intriguingly, USF2 has been implicated as a potential transcriptional modulator of Angiotensin II type 1 receptor (AT1R) - associated protein (ATRAP/Agtrap) in mice [19]. Altogether, the identified TFs are estimated to exhibit different mode of action depending on the CKD entity.

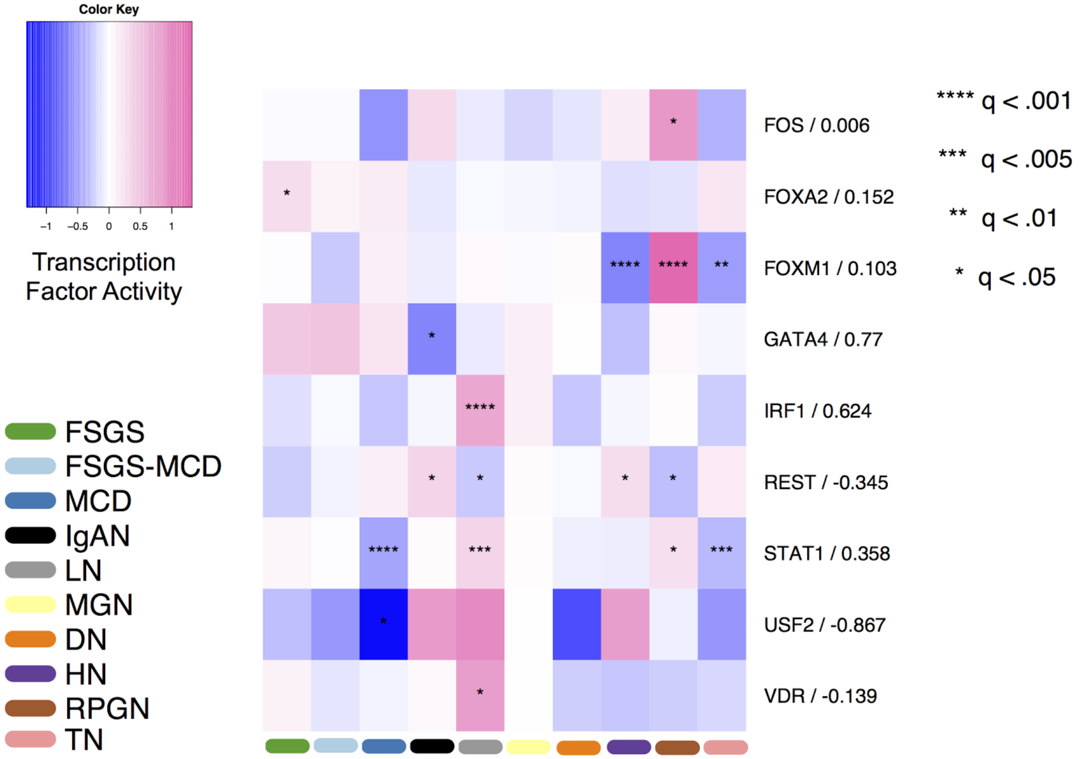


Figure 3. Transcription Factor Activity in glomerular CKD Entities. *Heatmap depicting transcription factor activity (colour) for each CKD entity and tumor nephrectomy in glomerular tissue. Negative numbers (blue) signify decreased transcription factor activity, positive numbers (pink) indicate increased transcription factor activity of an entity relative to the other entities. The corresponding q-value is represented by asterisk(s) (*) to indicate the statistical significance of each TF in each disease entity. The numbers to the right of factor names are Spearman's rank-based correlation coefficients of factor activity and factor expression across different CKD entities.*

2.4. Signaling Pathway Analysis

We complemented the functional characterization of transcription factor activities with an estimation of pathways activities. For this, we applied two methodologies, our tool PROGENy [20] and a general gene set enrichment tool, Piano [21].

2.4.1. Pathway activity of CKD entities using PROGENy

We first applied PROGENy, that infers pathway activity by looking at the changes in levels of the genes affected by perturbation on pathways. We have found that this provides a better proxy of pathway activity than looking at the genes in the actual pathway [20]. More specifically, whether it is expected to have a given PROGENy score for a particular pathway in a specific disease entity given the gene expression data at hand. Figure 4A depicts the pathway activities of CKD entities gained from PROGENy. Essentially, the degree of pathway deregulation is associated with the degree of disease severity. The pathways present in PROGENy exhibit rather divergent signaling footprints across the CKD entities. For example, VEGF is estimated to be significantly influential in five CKD entities: RPGN, HN, DN, LN and IgAN, from which, VEGF is predicted to be deactivated in RPGN and DN, but more prominently activated in HN, LN and IgAN. 10 out of 11 pathways are predicted to be significantly deregulated in RPGN with respect to TN, which is aligned with the diffusion map (Figure 2B) outcome; the divergence of RPGN from TN (control) is considerably prominent both at a global transcriptome landscape and signaling pathway level. Intriguingly the pathway JAK-STAT does not appear to be affected in RPGN, however, it is - JAK-STAT - considerably activated in LN and markedly deactivated in DN in comparison to TN. Overall, the separate CKD entities are characterised by distinct combinations, magnitudes and directions of signaling pathway activities according to PROGENy.

2.4.2. Pathway enrichment with Piano

While PROGENy can give accurate estimates of pathway activity, it is limited to 11 pathways for which robust signature could be generated [20]. To get a more global picture, we complemented that analysis with a gene-set-enrichment analysis using Piano [21]. A total of 160 pathways out of 1329 were significantly enriched (up-/down-regulated, corrected p-val < 0.05) in at least one CKD entity. Interestingly, no pathway was enriched in opposite directions across the CKD entities, in agreement with a common effect on kidney disease progression with a stereotypic tissue response regardless of the initial stimuli. HN was the entity with the largest number of differentially enriched pathways (81, 25 down-regulated, 56 up-regulated), while FSGS-MCD did not show significant enrichment for any pathway. Cell-cycle and immune-system related pathways were significantly up-regulated in seven out of the 9 different CKD entities (FSGS, HN, IgAN, LN, MGN and RPGN in both cases, DN for immune system, and MCD for cell-cycle); in contrast, VEGF pathway was differentially enriched in a single CKD entity (LN). Interestingly, TNFR2 pathway was differentially enriched in IgAN, HN, and LN, in line with the results from PROGENY where VEGF is significantly deregulated, not only in IgAN, HN and LN, but also in RPGN and DN. 59 different pathways showed significant enrichment in at least 3 CKD entities

(Figure 4B). Figure 4B also shows that HN (52), MGN (45), and IgAN (37) are the CKD entities with more pathways differentially enriched in at least 3 entities, result that agrees with Figure 2B showing these entities in the center of the diffusion map.

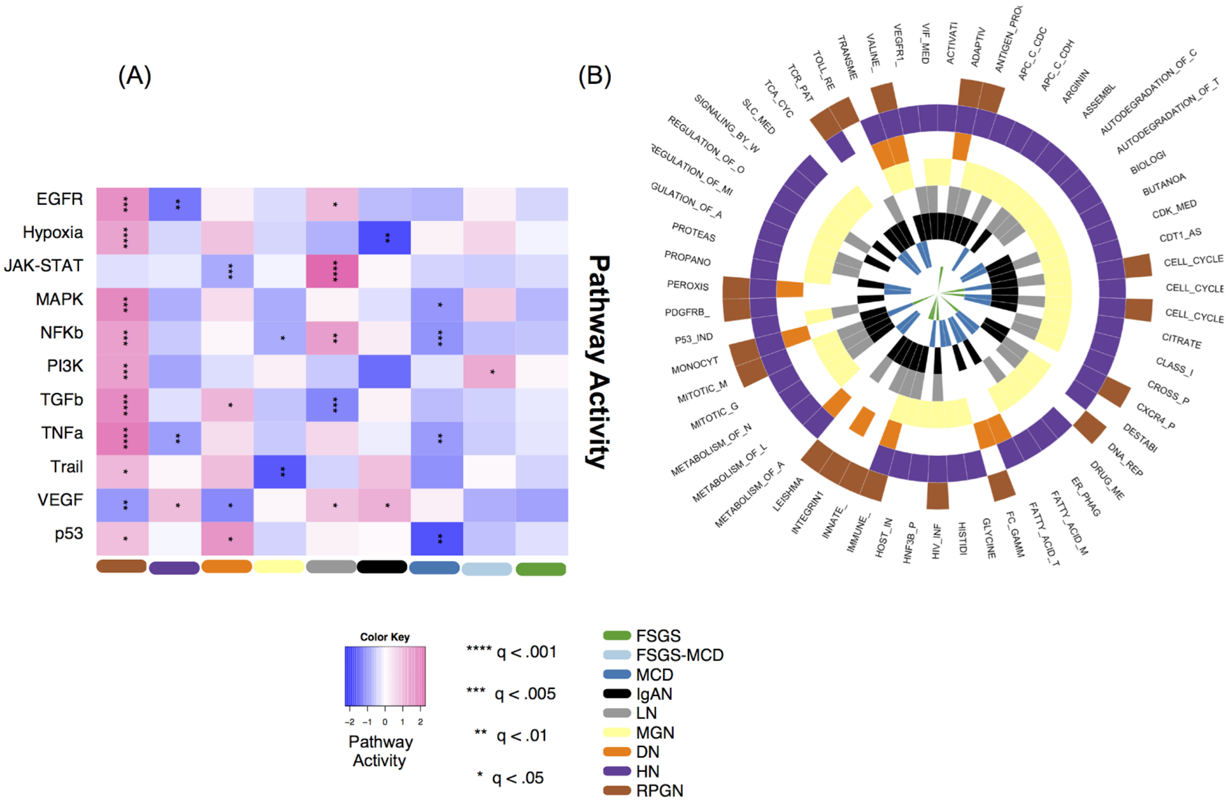
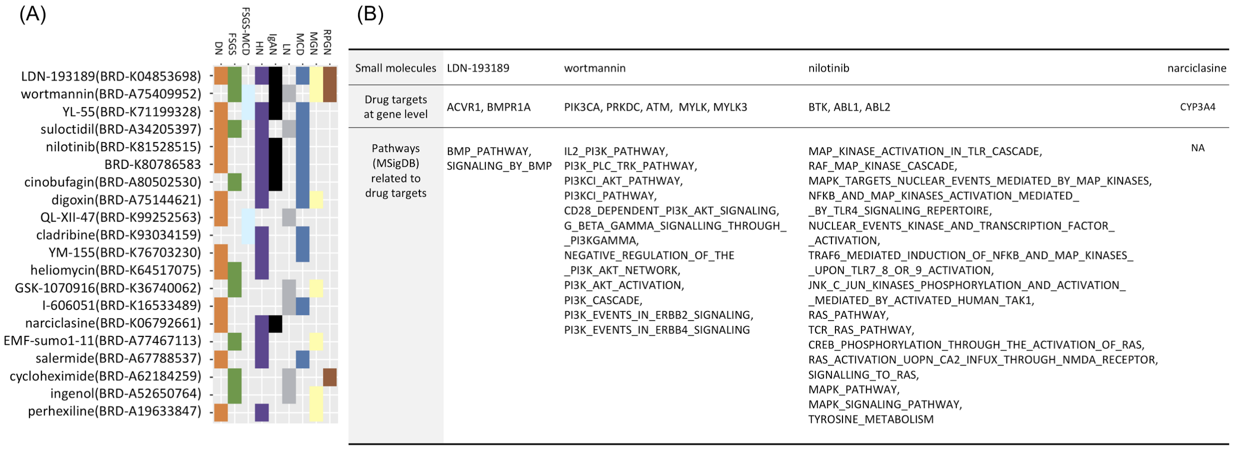


Figure 4 - Pathway activity alterations in CKD entities. *(A) Heatmap depicting pathway activity (colour) for each CKD entity relative to tumor nephrectomy in glomerular tissue, according to PROGENy [20]. The magnitude and direction - positive or negative - of PROGENy scores indicates the degree of pathway deregulation in a given CKD entity with regard to the reference condition, tumor nephrectomy. Permutation q-values are used to indicate statistical significance of each pathway in each disease entity, indicated by asterisk(s) (*). (B) Radial heatmap of consensually enriched pathways across three or more disease entities (up-, down-, or non-directional-regulation) according to PIANO [21] using MSigDB-C2-CP gene sets.*

2.5 Prediction of potential novel drugs indication to treat CKD

As a final analysis, we applied a signature-search-engine, L1000CDS^2 tool [22]. L1000CDS^2 measures the distance between two signatures of disease data and the LINCS-L1000 data, and then prioritizes small molecules that are expected to have reverse signature compared to disease signature. With L1000CDS^2, we performed this analysis separately for the nine CKD entities and then identified 220 small molecules across the CKD entitles (Sup. Figure 5). In order to narrow down the list of 220 small molecules, we focused on 20 small molecules observed in the L1000CDS^2 output of at least 3 subtypes (Figure 5A).



Small molecules	LDN-193189	wortmannin	nilotinib	narciclasine
Drug targets at gene level	ACVR1, BMPR1A	PIK3CA, PRKDC, ATM, MYLK, MYLK3	BTK, ABL1, ABL2	CYP3A4
Pathways (MSigDB) related to drug targets	BMP_PATHWAY, SIGNALING_BY_BMP	IL2_PI3K_PATHWAY, PI3K_PLC_TRK_PATHWAY, PI3KCI_AKT_PATHWAY, PI3KCI_PATHWAY, CD28_DEPENDENT_PI3K_AKT_SIGNALING, G_BETA_GAMMA_SIGNALLING_THROUGH__PI3KGAMMA, NEGATIVE_REGULATION_OF_THE_PI3K_AKT_NETWORK, PI3K_AKT_ACTIVATION, PI3K_CASCADE, PI3K_EVENTS_IN_ERBB2_SIGNALING, PI3K_EVENTS_IN_ERBB4_SIGNALING	MAP_KINASE_ACTIVATION_IN_TLR_CASCADE, RAF_MAP_KINASE_CASCADE, MAPK_TARGETS_NUCLEAR_EVENTS_MEDIATED_BY_MAP_KINASES, NFKB_AND_MAP_KINASES_ACTIVATION_MEDIATED__BY_TLR4_SIGNALING_REPERTOIRE, NUCLEAR_EVENTS_KINASE_AND_TRANSCRIPTION_FACTOR__ACTIVATION, TRAF6_MEDIATED_INDUCTION_OF_NFKB_AND_MAP_KINASES__UPON_TLR7_8_OR_9_ACTIVATION, JNK_C_JUN_KINASES_PHOSPHORYLATION_AND_ACTIVATION__MEDIATED_BY_ACTIVATED_HUMAN_TAK1, RAS_PATHWAY, TCR_RAS_PATHWAY, CREB_PHOSPHORYLATION_THROUGH_THE_ACTIVATION_OF_RAS, RAS_ACTIVATION_UOPN_CA2_INFUX_THROUGH_NMDA_RECEPTOR, SIGNALLING_TO_RAS, MAPK_PATHWAY, MAPK_SIGNALING_PATHWAY, TYROSINE_METABOLISM	NA

Figure 5. - Top 20 Drug Candidates from Drug Repositioning. *(A) Distribution of 20 small molecules reversely correlated with at least 3 CKD entities. (B) Table of four small molecules out of the 20 of (A) supported by manual curation. Table shows drugs (first row), protein coding genes targeted by these four drugs (second row) and pathways (MSigDB) related to the biological functions these drugs affect (third row).*

By manual curation of scientific publications, we found that four small molecules have experimental evidence to support their clinical relevance in CKD or renal disease animal model testing (Sup. Table 1). BRD-K04853698 (LDN-193189) which is known as a selective BMP signaling inhibitor, and has been shown to suppress endothelial damage in mice with chronic kidney disease [23]. Wortmannin is one of cell-permeable PI3K inhibitors, and it has been shown to decreases albuminuria and recovers podocyte damage for early diabetic nephropathy in rat[24]. The tyrosine kinase inhibitor Nilotinib is a Food and Drug Administration (FDA) approved medication to treat chronic myelogenous leukemia (CML).[25] Iyoda et al. showed [26] that nilotinib treatment resulted in stabilized kidney function and prolonged survival after subtotal nephrectomy in rats when compared to vehicle treatment b [26]. Finally narciclasine was identified which has been reported to reduce macrophage infiltration and inflammation in the mouse unilateral ureteral obstruction (UUO) model of kidney fibrosis [27].

To further explore the association of these drugs with CKD disease and progression, we analysed the expression data for the targets of the literature supported drug candidates. First, each drug candidate was mapped to genes that encode the proteins targeted by these drugs (Figure 5B). For each gene, its differential expression of any CKD entity against TN was evaluated. Out of the 11 mapped genes, MYLK3, a target of narciclasine, was significantly differentially expressed (under-expressed, logFC<-1, $p<0.05$) in two CKD entities (IgAN and LN) (Supplementary figure 6). Complementarily, screened drugs were mapped to the pathways they affect based on their functional information. The enrichment of the subset of pathways was evaluated using the previous results from gene set analysis algorithm (piano). This time, only PI3KCI pathway appears to be both, significantly enriched for HN patient data (up-regulated, $p<0.05$), and as pathway affected by the candidate repositioned drugs (Wortmannin, PI3K inhibitor).

3. Discussion

In this paper, we have aimed to shed light on the commonalities and differences among glomerular transcriptomics of major kidney diseases contributing to the CKD epidemic affecting >10% of the population in Europe and the United States. Multiple pathologies are covered under the broad umbrella of being contributors to CKD and, while they share a physiological manifestation in terms of loss of kidney function, the driving molecular process can be very different. In this study we explored these processes by analyzing glomerular gene expression data from kidney biopsies obtained upon microdissection. We found general trends such as underexpression of SLC6A8, ALB and CALB1 consistently across all included kidney disease entities. The decreased expression of SLC6A8 which encodes for a creatinine transporter might just reflect a negative feedback loop due to increased creatinine levels with kidney functional decline across all disease entities. Similarly, although albumin is primarily expressed in the liver, there are reports about minor albumin expression in the kidney[28] and thus decreased expression of the gene encoding for albumin (ALB) might be a negative feedback loop of cells exposed to high urinary albumin levels during progression of CKD. CALB1 encodes for calbindin 1 an intracellular calcium binding protein which has been reported to be downregulated in the rat UUO kidney-fibrosis model and the anti-glomerular basement membrane glomerulonephritis model [29]. Decreased urinary calbindin 1 was proposed as a biomarker for kidney injury [29].

Other genes were specifically altered in certain kidney disease entities such as Quaking (QKI) or Lysozyme C (LYZ), significantly overexpressed/underexpressed/non-significant depending on the underlying kidney disease. It is known that QKI is associated with angiogenic growth factor release and plays a pathological role in the kidney [30], while LYZ was known to be related with vascular damage and heart failure but was recently found to be increased in plasma during CKD progression [31]. This data supports the fact that despite a stereotypic response of the kidney to injury with glomerulosclerosis, interstitial fibrosis and nephron loss there are various disease specific differences that are important to understand in order to develop novel personalized therapeutics.

CKD is a complex disease with a high degree of polygenicity. Furthermore, it is a very heterogeneous condition that can be acquired through a variety of biological mechanisms which is reflected by the results of pathway analysis. There was little to no overlap in significantly enriched pathways between the different kidney disease entities. We found 59 different pathways that showed significant enrichment in at least 3 disease entities (Figure 4B), indicating that different disease entities share some general mechanisms but their underlying pathophysiology differs from one entity to another. Besides increasing the interpretability, the pathway analysis identify many more differences among disease-identities than the gene-level analysis (Figure 2A). For example, pathway analysis identified pathways related to the metabolism of lipids and lipoproteins significantly down-regulated in MCD, MGN, and HN; and pathways related to fatty acid metabolism significantly down-regulated in MCD, IgAN, MGN, and HN, results similar to those reported by Kang et al [6].

PROGENy (Figure 4A) estimated JAK-STAT, a major cytokine signal transduction regulator [32], to be significantly activated in LN with respect to TN and DoROthEA (Figure 3) predicted the TFs IRF1 and STAT1 to be significantly enriched in LN. The estimated pathogenic influentiality of JAK-STAT/STAT1/Interferon signaling in LN is supported by various studies [33] [34] [35].

We also used the signature-matching paradigm to explore potential drugs that could revert the disease phenotype, and found that four drugs could be an encouraging medicine for CKD entities. Even though more experimental validations are required for the unknown medical interaction between drugs of our results and CKD progression, our approach suggests that it is possible to find promising treatments for CKD with the concept of drug repositioning. In particular, for one of the identified drugs, nilotinib, results have already been confirmed that it is safe to use for treatment and there is supporting data of its value insight at indications for CKD [26].

The analysis of the drug targets' expression found that MYLK3, a gene encoding for one of the targets of Narciclasine, was significantly underexpressed in IgAN and LN when compared with TN. Similarly, PI3KCI pathway, the target of Wortmannin was enriched in HN (up-regulated, $p<0.05$). This analysis attempts to refine the outcome of the repositioning analysis, and at the same time help to connect it to the disease mechanism both in gene as well as in pathway level.

We see our analysis as a first step towards a characterization of the similarities and differences of the various pathologies that lead to CKD. As more data sets become available, either micro-array or RNA-seq, these can be integrated in our pipeline. Furthermore, the burgeoning field of single-cell RNA (scRNA) has started to produce data sets in kidney [36,37], which hold the potential to revolutionize our understanding of the functioning of the kidney and its pathologies [38] [39]. In particular, scRNA data can provide us signatures of the many cell types of the kidney, which in turn can be used to deconvolute the composition of cell types[12] in the more abundant and cost-effective bulk expression datasets [39]. Other data sets, such as (phospho)proteomics[40] and metabolomics[41], would complement gene expression towards a more complete picture of the CKD-entities. Ideally, all these data sets would be collected in a standardized manner to facilitate integration, which was a major hurdle in our study. Such a comprehensive analysis across large cohorts, akin to what has happened for the different tumour types thanks to initiatives such as the International Cancer Genome Consortium, can lead to major improvements in our understanding of and treatment venues for CKD [42].

4. Methods

4.1. Data collection

Raw data CEL files of each microarray dataset - GSE20602 [10]; GSE32591 [11]; GSE37460 [11]; GSE47183 [12,13]; GSE50469 [14] - were downloaded and imported to R (R version 3.3.2) using the getGEOSuppFiles and read.affy function of the GEOquery and simpleaffy package, respectively [43]. Each dataset came from either Affymetrix Human Genome U133A Array or Affymetrix Human Genome U133 Plus2.0 array, therefore the preprocessing was done with the affy R package [44] accordingly.

4.2. Preprocessing and mapping

RNA quality was assessed by RNA degradation plots using the AffyRNAdeg function from the affy package. In order to assess the statistical characteristics of the raw data, the affyPLM package [45] was used and probe-level metric calculations were carried out on the CEL files by calling the fitPLM function. The homogeneity of probe sets was evaluated by Normalized Unscaled Standard Error (NUSE) and Relative Log Expression (RLE) boxplots [46–48]. We removed all arrays that showed greater spread of NUSE value distribution with respect to the rest or where the median NUSE value was above 1.05, as these features indicated the sign of low quality array. The RLE values represent the ratio between the expression of a probe set and the median expression of that probe set across all arrays in the data set. The ratios are expected to be centered around zero on a logarithmic scale. RLE boxplots were generated to visualise the distribution of RLE values. Array quality was evaluated by taking NUSE and RLE plots into account.

The preprocessing step also constituted background correction and log2 transformation of the raw values, of which was done by the Robust Multichip Average (RMA) package [49–51].

Probe IDs were mapped to Entrez Gene ID resulting in 12437 (Platform GPL570, Affymetrix Human Genome U133 Plus 2.0 Array) and 20514 (Platform GPL96, Affymetrix Human Genome U133A Array) unique Entrez gene identifiers, respectively. In the case where datasets contained multiple probes for the same Entrez ID gene, the probe with the highest interquartile range (IQR) was retained as the representative of that given gene in the dataset. For this filtering step, the nsFilter function from the genefilter package [52] was utilized.

4.3. Correlation of arrays

Correlation of arrays was assessed by hierarchical clustering of the arrays based on gene expression Spearman's rank-based correlation coefficients. Low Spearman correlation coefficients imply considerable differences between array intensities [53].

4.4. Normalization and batch effect mitigation

When required, cyclic loess normalization was applied using the limma package [16,54,55] or YuGene transformation was carried out using the YuGene R package [56].

The efficient integration of the data from different sources and platforms requires batch effect management, which should be customised to the data at hand. The current data was heavily affected by platform- and study-specific batch effects, because the outcome categories (CKD entities and their samples) were unevenly distributed across the batches. The commonly used algorithms for correcting batch effects assume a balanced distribution of outcome categories across batches and are vulnerable to the group-batch imbalance [57–60]. We conducted a stringent batch effect mitigation process in order to minimize the influence of technical heterogeneity. First, we structured the data in a platform-specific manner. Then, we conducted differential gene expression analysis between those identical biological conditions that are originating from distinct study sources after cyclic loess normalization and removed those genes that are significantly differentially expressed between them, as it indicated difference mainly due to the data source, rather than the biological difference. We applied this procedure for the data fragments coming from Affymetrix Human Genome U133 Plus 2.0 Array and Affymetrix Human Genome U133A Array. Next, we merged the data sets between the two platforms using the overlapping genes, followed by a process to mitigate the platform-induced batch effect. This latter procedure is similar to the one used for the data source-specific batch effect mitigation. By applying this stringent procedure, we eliminated the genes that are the most affected by batch effects. For the illustration of this procedure, see Sup. Figure 1. The scater R package [61] was used for producing the batch effect management related plots.

4.5. Detection of genes with consistently small p-values across all studies

Based on the assumption that common mechanisms might contribute to all CKD entities we performed a Maximum p-value (maxP) method [62] - which uses the maximum p-value as the test statistic - on the output of the differential expression analysis of the hypothetically separate studies. The maxP test follows a beta distribution that is parametrized by $\alpha = K$ and $\beta = 1$ under the null hypothesis:

$H_0 : \cap_{k=1}^{K} \{\theta k = 0\}$ versus $H_a : \cap_{k=1}^{K} \{\theta k \neq 0\}$ (HS_A)

where θ_k is the effect size of study k.

This hypothesis setting (symbolised by HS_A) aims to uncover differentially expressed (DE) genes that acquire non-zero effect sizes in all studies. To phrase it differently, it is designated to unravel DE genes that are characterised by a small p-value across all studies [62–64].

To obtain the p-values, differential expression analysis was conducted on the batch effect mitigated data using the limma R package [15]. We contrasted each glomerular CKD entity with tumor nephrectomy condition - each CKD - tumor nephrectomy contrast represented a hypothetically separate "study" - and the lmFit function was used to fit a linear model to the expression data for each probe set in the array series, followed by the estimation of eBayes values and the execution of a moderated t-test by the empirical Bayes method for differential expression (eBayes function) [15,53].

4.6. Diffusion map

The batch mitigated data containing merely the maxP identified (section 4.5.) 1790 genes (FDR < 0.01) (Sup. file 1), were YuGene transformed [56] and the destiny R package [65] was utilised to produce the diffusion maps.

4.7. Functional Analysis

4.7.1. Transcription factor activity analysis

We estimated transcription factor activities in the glomerular CKD entities using DoRothEA[17] which is a pipeline that tries to estimate transcription factor activity via the expression level of its target genes utilizing a curated database of transcription factor - target gene interactions (TF Regulon). The cyclic loess normalized expression values of all genes in all conditions were scaled and re-centered across the conditions and the transcription factors activities were estimated from the TF Regulon using VIPER [18]. We then conducted a Spearman's rank-based correlation between the identified transcription factors' activity and the scaled and re-centered expression of the genes encoding for these transcription factors. Since as a consequence of the batch effect mitigation procedure we lost many potentially informative genes, the coverage of the TF regulon database was limited and hence our statistical power decrease, meaning that there might be more differentially regulated TFs.

4.7.2. Inferring Signaling Pathway Activity fusing PROGENy

We used the cyclic loess normalised and batch effect mitigated expression values for PROGENy [20], a method which utilizes downstream gene expression changes due to pathway perturbation in order to infer the upstream signaling pathway activity. The expression values were standardised to express a distance of each CKD sample from tumor nephrectomy (CKD entity sample scaled by the standard deviation of tumor nephrectomy). We used the overlapping genes between the standardised gene expression matrix and the PROGENy model. Then, a matrix multiplication was done to get the product matrix, containing a PROGENy score for each pathway in each CKD entity. A positive PROGENy score in a given pathway in a given CKD entity implies higher signaling activity compared to that specific pathways' activity in tumor nephrectomy, and vice versa for a negative PROGENy score.

Statistical significance was assessed using permutation-based hypothesis testing. We resampled the standardised gene expression values in a way that results in the randomised allocation of expression values to different glomerular disease labels. This resampling was done ten thousand times. We then computed PROGENy scores from these permuted Z-scores, resulting in a list of glomerular CKD entity specific PROGENy scores. By applying this approach we generated an empirical null distribution on the basis of the original gene expression sample distribution. The probability that the original PROGENy score in a given glomerular CKD entity is coming from the estimated null distribution or not was evaluated in a pathway-specific manner. We used a p-value of 0.05 as the threshold for statistical significance. Furthermore, we applied the Benjamini-Hochberg adjustment [66] on the p-values to correct for multiple testing.

4.7.3. Pathway Analysis with Piano

Pathway analysis was performed using the piano package from R [21]. The Molecular Signature Database - Curated Pathways - Canonical Pathways (MSigDB-C2-CP) was used as biological model to map the individual genes to functional sets. Gene-level statistics were obtained after applying the limma algorithm (see section 4.5.). All disease entities were compared to tumor nephrectomy, because the healthy living donor samples were highly corrupted by batch effects and as a result of the batch effect mitigation, we had to remove a considerably large number of genes from these samples. The following ten methods (with their corresponding gene-level statistics) were used as input of the pathway analysis algorithm to calculate gene set enrichment: Fisher (PVal), Stouffer (PVal), Reporter (PVal), PAGE (TVal), Tail Strength (PVal) ,GSEA (TVal), Mean (FC), Median (FC), Sum (FC), MaxMean (TVal). From algorithm's output only the adjusted p-values from p_distinct_up, p_non_dir, and p_mix_down were extracted. For each pathway/p-value pair the geometrical average across all ten methods was calculated.

4.8. Drug repositioning

Cyclic loess normalized gene expression data for nine glomerular CKD entities were analyzed separately for measuring characteristic direction (CD) [67]. Cosine distance for each gene was computed to the line which has 90 degree to the hyperplane which set the given CKD entity apart from tumor nephrectomy in N-dimensional gene expression space. Then, for each CKD entity, the signature of cosine distances computed by characteristic direction was applied to a signature-search-engine, L1000CDS^2 [22] with the mode of reverse in configuration. L1000CDS^2 provided the top 50 ranked small molecule candidates with 1-cos(a), p-value, drug database links. Significant small molecules with FDR < 0.05 were filtered in for the nine CKD entities, separately. For converting the name of small molecules into general chemical names, we referred to LINCS phase I, II dataset stored in GEO (GSE92742, GSE70138) [68].

Acknowledgements

This work was supported by the European Union Horizon 2020 grant SyMBioSys MSCA-ITN-2015-ETN #675585 that provided the financial support for A.A.

References

1. Hamer RA, El Nahas AM: The burden of chronic kidney disease. *BMJ* 332: 563–564, 2006

2. American Society of Nephrology: Kidney Research is Underfunded Compared to the Cost of Care. Available from: www.asn-online.org/policy/webdocs/KidneyResearch.pdf

3. Beckerman P, Qiu C, Park J, Ledo N, Ko Y-A, Park A-SD, Han S-Y, Choi P, Palmer M, Susztak K: Human Kidney Tubule-Specific Gene Expression Based Dissection of Chronic Kidney Disease Traits. *EBioMedicine* 24: 267–276, 2017

4. Nair V, Komorowsky CV, Weil EJ, Yee B, Hodgin J, Harder JL, Godfrey B, Ju W, Boustany-Kari CM, Schwarz M, Lemley KV, Nelson PJ, Nelson RG, Kretzler M: A molecular morphometric approach to diabetic kidney disease can link structure to function and outcome. *Kidney Int.* [Internet] 2017 Available from: http://dx.doi.org/10.1016/j.kint.2017.08.013

5. Schena FP, Nistor I, Curci C: Transcriptomics in kidney biopsy is an untapped resource for precision

therapy in nephrology: a systematic review. *Nephrol. Dial. Transplant* [Internet] 2017 Available from: https://academic.oup.com/ndt/article/doi/10.1093/ndt/gfx211/4057587/Transcriptomics-in-kidney-biopsy-is-an-untapped [cited 2017 Sep 14]

6. Kang HM, Ahn SH, Choi P, Ko Y-A, Han SH, Chinga F, Park ASD, Tao J, Sharma K, Pullman J, Bottinger EP, Goldberg IJ, Susztak K: Defective fatty acid oxidation in renal tubular epithelial cells has a key role in kidney fibrosis development. *Nat. Med.* 21: 37–46, 2015

7. Ju W, Nair V, Smith S, Zhu L, Shedden K, Song PXK, Mariani LH, Eichinger FH, Berthier CC, Randolph A, Lai JY-C, Zhou Y, Hawkins JJ, Bitzer M, Sampson MG, Thier M, Solier C, Duran-Pacheco GC, Duchateau-Nguyen G, Essioux L, Schott B, Formentini I, Magnone MC, Bobadilla M, Cohen CD, Bagnasco SM, Barisoni L, Lv J, Zhang H, Wang H-Y, Brosius FC, Gadegbeku CA, Kretzler M, ERCB, C-PROBE, NEPTUNE, and PKU-IgAN Consortium: Tissue transcriptome-driven identification of epidermal growth factor as a chronic kidney disease biomarker. *Sci. Transl. Med.* 7: 316ra193, 2015

8. Edgar R, Domrachev M, Lash AE: Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* 30: 207–210, 2002

9. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A, Lee H, Zhang N, Robertson CL, Serova N, Davis S, Soboleva A: NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Res.* 41: D991–5, 2013

10. Neusser MA, Lindenmeyer MT, Moll AG, Segerer S, Edenhofer I, Sen K, Stiehl DP, Kretzler M, Gröne H-J, Schlöndorff D, Cohen CD: Human nephrosclerosis triggers a hypoxia-related glomerulopathy. *Am. J. Pathol.* 176: 594–607, 2010

11. Berthier CC, Bethunaickan R, Gonzalez-Rivera T, Nair V, Ramanujam M, Zhang W, Bottinger EP, Segerer S, Lindenmeyer M, Cohen CD, Davidson A, Kretzler M: Cross-species transcriptional network analysis defines shared inflammatory responses in murine and human lupus nephritis. *J. Immunol.* 189: 988–1001, 2012

12. Ju W, Greene CS, Eichinger F, Nair V, Hodgin JB, Bitzer M, Lee Y-S, Zhu Q, Kehata M, Li M, Jiang S, Rastaldi MP, Cohen CD, Troyanskaya OG, Kretzler M: Defining cell-type specificity at the transcriptional level in human disease. *Genome Res.* 23: 1862–1873, 2013

13. Martini S, Nair V, Keller BJ, Eichinger F, Hawkins JJ, Randolph A, Böger CA, Gadegbeku CA, Fox CS, Cohen CD, Kretzler M, European Renal cDNA Bank, C-PROBE Cohort, CKDGen Consortium: Integrative biology identifies shared transcriptional networks in CKD. *J. Am. Soc. Nephrol.* 25: 2559–2572, 2014

14. Hodgin JB, Berthier CC, John R, Grone E, Porubsky S, Gröne H-J, Herzenberg AM, Scholey JW, Hladunewich M, Cattran DC, Kretzler M, Reich HN: The molecular phenotype of endocapillary proliferation: novel therapeutic targets for IgA nephropathy. *PLoS One* 9: e103413, 2014

15. Phipson B, Lee S, Majewski IJ, Alexander WS, Smyth GK: ROBUST HYPERPARAMETER ESTIMATION PROTECTS AGAINST HYPERVARIABLE GENES AND IMPROVES POWER TO DETECT DIFFERENTIAL EXPRESSION. *Ann. Appl. Stat.* 10: 946–963, 2016

16. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK: limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43: e47, 2015

17. Garcia-Alonso LM, Iorio F, Matchan A, Fonseca NA, Jaaks P, Peat G, Pignatelli M, Falcone F, Benes CH, Dunham I, Bignell GR, McDade S, Garnett MJ, Saez-Rodriguez J: Transcription factor activities enhance markers of drug sensitivity in cancer. *Cancer Res.* [Internet] 2017 Available from: http://dx.doi.org/10.1158/0008-5472.CAN-17-1679

18. Alvarez MJ, Shen Y, Giorgi FM, Lachmann A, Belinda Ding B, Hilda Ye B, Califano A: Functional

characterization of somatic mutations in cancer using network-based inference of protein activity. *Nat. Genet.* 48: 838–847, 2016

19. Matsuda M, Tamura K, Wakui H, Maeda A, Ohsawa M, Kanaoka T, Azushima K, Uneda K, Haku S, Tsurumi-Ikeya Y, Toya Y, Maeshima Y, Yamashita A, Umemura S: Upstream stimulatory factors 1 and 2 mediate the transcription of angiotensin II binding and inhibitory protein. *J. Biol. Chem.* 288: 19238–19249, 2013

20. Schubert M, Klinger B, Klünemann M, Sieber A, Uhlitz F, Sauer S, Garnett MJ, Blüthgen N, Saez-Rodriguez J: Perturbation-response genes reveal signaling footprints in cancer gene expression. *Nat. Commun.* 9: 20, 2018

21. Väremo L, Nielsen J, Nookaew I: Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods. *Nucleic Acids Res.* 41: 4378–4391, 2013

22. Duan Q, Reid SP, Clark NR, Wang Z, Fernandez NF, Rouillard AD, Readhead B, Tritsch SR, Hodos R, Hafner M, Niepel M, Sorger PK, Dudley JT, Bavari S, Panchal RG, Ma'ayan A: L1000CDS2: LINCS L1000 characteristic direction signatures search engine. *NPJ Syst Biol Appl* [Internet] 2: 2016 Available from: http://dx.doi.org/10.1038/npjsba.2016.15

23. Kajimoto H, Kai H, Aoki H, Uchiwa H, Aoki Y, Yasuoka S, Anegawa T, Mishina Y, Suzuki A, Fukumoto Y, Imaizumi T: BMP type I receptor inhibition attenuates endothelial dysfunction in mice with chronic kidney disease. *Kidney Int.* 87: 128–136, 2015

24. Kim SH, Jang YW, Hwang P, Kim HJ, Han GY, Kim CW: The reno-protective effect of a phosphoinositide 3-kinase inhibitor wortmannin on streptozotocin-induced proteinuric renal disease rats. *Exp. Mol. Med.* 44: 45–51, 2012

25. Blay J-Y, von Mehren M: Nilotinib: a novel, selective tyrosine kinase inhibitor. *Semin. Oncol.* 38 Suppl 1: S3–9, 2011

26. Iyoda M, Shibata T, Hirai Y, Kuno Y, Akizawa T: Nilotinib attenuates renal injury and prolongs survival in chronic kidney disease. *J. Am. Soc. Nephrol.* 22: 1486–1496, 2011

27. Fuchs S, Hsieh LT, Saarberg W, Erdelmeier CAJ, Wichelhaus TA, Schaefer L, Koch E, Fürst R: Haemanthus coccineusextract and its main bioactive component narciclasine display profound anti-inflammatory activitiesin vitroandin vivo. *J. Cell. Mol. Med.* 19: 1021–1032, 2015

28. Nahon JL, Tratner I, Poliard A, Presse F, Poiret M, Gal A, Sala-Trepat JM, Legrès L, Feldmann G, Bernuau D: Albumin and alpha-fetoprotein gene expression in various nonhepatic rat tissues. *J. Biol. Chem.* 263: 11436–11442, 1988

29. Iida T, Fujinaka H, Xu B, Zhang Y, Magdeldin S, Nameta M, Liu Z, Yoshida Y, Yaoita E, Tomizawa S, Saito A, Yamamoto T: Decreased urinary calbindin 1 levels in proteinuric rats and humans with distal nephron segment injuries. *Clin. Exp. Nephrol.* 18: 432–443, 2014

30. Gomez IG, Nakagawa N, Duffield JS: MicroRNAs as novel therapeutic targets to treat kidney injury and fibrosis. *Am. J. Physiol. Renal Physiol.* 310: F931–44, 2016

31. Glorieux G, Mullen W, Duranton F, Filip S, Gayrard N, Husi H, Schepers E, Neirynck N, Schanstra JP, Jankowski J, Mischak H, Argilés À, Vanholder R, Vlahou A, Klein J: New insights in molecular mechanisms involved in chronic kidney disease using high-resolution plasma proteome analysis. *Nephrol. Dial. Transplant* 30: 1842–1852, 2015

32. Rawlings JS, Rosler KM, Harrison DA: The JAK/STAT signaling pathway. *J. Cell Sci.* 117: 1281–1283, 2004

33. Dong J, Wang QX, Zhou CY, Ma XF, Zhang YC: Activation of the STAT1 signalling pathway in lupus nephritis in MRL/lpr mice. *Lupus* 16: 101–109, 2007

34. Wang S, Yang N, Zhang L, Huang B, Tan H, Liang Y, Li Y, Yu X: Jak/STAT signaling is involved in the inflammatory infiltration of the kidneys in MRL/lpr mice. *Lupus* 19: 1171–1180, 2010

35. Ripoll È, de Ramon L, Draibe Bordignon J, Merino A, Bolaños N, Goma M, Cruzado JM, Grinyó JM, Torras J: JAK3-STAT pathway blocking benefits in experimental lupus nephritis. *Arthritis Res. Ther.* 18: 134, 2016

36. Sivakamasundari V, Bolisetty M, Sivajothi S: Comprehensive Cell Type Specific Transcriptomics of the Human Kidney. *bioRxiv* [Internet] 2017 Available from: https://www.biorxiv.org/content/early/2017/12/21/238063.abstract

37. Wu H, Uchimura K, Donnelly E, Kirita Y, Morris SA: Comparative analysis of kidney organoid and adult human kidney single cell and single nucleus transcriptomes. *bioRxiv* [Internet] 2017 Available from: https://www.biorxiv.org/content/early/2017/12/11/232561.abstract

38. Wu H, Humphreys BD: The promise of single-cell RNA sequencing for kidney disease investigation. *Kidney Int.* 92: 1334–1342, 2017

39. Kiryluk K, Bomback AS, Cheng Y-L, Xu K, Camara PG, Rabadan R, Sims PA, Barasch J: Precision Medicine for Acute Kidney Injury (AKI): Redefining AKI by Agnostic Kidney Tissue Interrogation and Genetics. *Semin. Nephrol.* 38: 40–51, 2018

40. Liu P, Lassén E, Nair V, Berthier CC, Suguro M, Sihlbom C, Kretzler M, Betsholtz C, Haraldsson B, Ju W, Ebefors K, Nyström J: Transcriptomic and Proteomic Profiling Provides Insight into Mesangial Cell Function in IgA Nephropathy. *J. Am. Soc. Nephrol.* 28: 2961–2972, 2017

41. Hocher B, Adamski J: Metabolomics for clinical use and research in chronic kidney disease. *Nat. Rev. Nephrol.* 13: 269–284, 2017

42. Mariani LH, Pendergraft WF 3rd, Kretzler M: Defining Glomerular Disease in Mechanistic Terms: Implementing an Integrative Biology Approach in Nephrology. *Clin. J. Am. Soc. Nephrol.* 11: 2054–2060, 2016

43. Wilson CL, Miller CJ: Simpleaffy: a BioConductor package for Affymetrix Quality Control and data analysis. *Bioinformatics* 21: 3683–3685, 2005

44. Gautier L, Cope L, Bolstad BM, Irizarry RA: affy--analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* 20: 307–315, 2004

45. Bolstad BM: affyPLM: Methods for fitting probe-level models. *BioConductor version 2. 0 package* 2007

46. Bolstad BM: Low-level Analysis of High-density Oligonucleotide Array Data: Background, Normalization and Summarization. 2004. *University of California, Berkeley.(http://bmbolstad. com*

47. Bolstad BM, Collin F, Brettschneider J, Simpson K, Cope L, Irizarry RA, Speed TP: Quality Assessment of Affymetrix GeneChip Data. In: *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, pp 33–47, 2005

48. Brettschneider J, Collin F, Bolstad BM, Speed TP: Quality Assessment for Short Oligonucleotide Microarray Data. *Technometrics* 50: 241–264, 2008

49. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP: Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.* 31: e15, 2003

50. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP: Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4: 249–264, 2003

51. Bolstad BM, Irizarry RA, Astrand M, Speed TP: A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 19: 185–193, 2003

52. Gentleman R, Carey V, Huber W, Hahne F: Genefilter: Methods for filtering genes from microarray experiments. *R package version* [Internet] 2011 Available from: ftp://ftp2.uib.no/pub/bioconductor/2.7/bioc/html/genefilter.html

53. Baetke SC, Adriaens ME, Seigneuric R, Evelo CT, Eijssen LMT: Molecular pathways involved in prostate carcinogenesis: insights from public microarray datasets. *PLoS One* 7: e49831, 2012

54. Smyth GK: limma: Linear Models for Microarray Data. In: *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, pp 397–420, 2005

55. Smyth GK, Speed T: Normalization of cDNA microarray data. *Methods* 31: 265–273, 2003

56. Lê Cao K-A, Rohart F, McHugh L, Korn O, Wells CA: YuGene: a simple approach to scale gene expression data derived from different platforms for integrated analyses. *Genomics* 103: 239–251, 2014

57. Johnson WE, Li C, Rabinovic A: Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 8: 118–127, 2007

58. Goh WWB, Wang W, Wong L: Why Batch Effects Matter in Omics Data, and How to Avoid Them. *Trends Biotechnol.* 35: 498–507, 2017

59. Nygaard V, Rødland EA, Hovig E: Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. *Biostatistics* 17: 29–39, 2016

60. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K, Irizarry RA: Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* 11: 733, 2010

61. McCarthy DJ, Campbell KR, Lun ATL, Wills QF: Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics* 33: 1179–1186, 2017

62. Wilkinson B: A statistical consideration in psychological research. *Psychol. Bull.* 48: 156–158, 1951

63. Chang L-C, Lin H-M, Sibille E, Tseng GC: Meta-analysis methods for combining multiple expression profiles: comparisons, statistical characterization and an application guideline. *BMC Bioinformatics* 14: 368, 2013

64. Song C, Tseng GC: HYPOTHESIS SETTING AND ORDER STATISTIC FOR ROBUST GENOMIC META-ANALYSIS. *Ann. Appl. Stat.* 8: 777–800, 2014

65. Angerer P, Haghverdi L, Büttner M, Theis FJ, Marr C, Buettner F: destiny: diffusion maps for large-scale single-cell data in R. *Bioinformatics* 32: 1241–1243, 2016

66. Benjamini Y, Hochberg Y: Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Series B Stat. Methodol.* 57: 289–300, 1995

67. Clark NR, Hu KS, Feldmann AS, Kou Y, Chen EY, Duan Q, Ma'ayan A: The characteristic direction: a geometrical approach to identify differentially expressed genes. *BMC Bioinformatics* 15: 79, 2014

68. Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X, Gould J, Davis JF, Tubelli AA, Asiedu JK, Lahr DL, Hirschman JE, Liu Z, Donahue M, Julian B, Khan M, Wadden D, Smith IC, Lam D, Liberzon A, Toder C, Bagul M, Orzechowski M, Enache OM, Piccioni F, Johnson SA, Lyons NJ, Berger AH, Shamji AF, Brooks AN, Vrcic A, Flynn C, Rosains J, Takeda DY, Hu R, Davison D, Lamb J, Ardlie K, Hogstrom L, Greenside P, Gray NS, Clemons PA, Silver S, Wu X, Zhao W-N, Read-Button W, Wu X, Haggarty SJ, Ronco LV, Boehm JS, Schreiber SL, Doench JG, Bittker JA, Root DE, Wong B, Golub TR: A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell* 171: 1437–1452.e17, 2017

Supplementary material

Illustration of the batch effect mitigation procedure

For illustrative purposes the batch effect mitigation process will be illuminated using samples from the condition IgA Nephropathy (IgAN). IgAN samples are only represented in platform Affymetrix Human Genome U133A Array (GPL96), however, these samples are originating from two distinct studies, GSE37460 [11] and GSE50469 [14], respectively. Sup. Figure 1A depicts a Principal component analysis (PCA) of gene expression measurements corresponding to the aforementioned IgAN samples from the two distinct studies prior batch effect mitigation. The samples are colored according to their respective study origin. Based on the PCA plot it can be said that there is a considerable clustering of samples due to the different study origin. In other words, the Principal component 2 explains the variance induced by the study source dissimilarity. We conducted differential expression analysis using limma [15] between the IgAN samples of GSE37460 and that of GSE50469. Sup. Figure 1B shows an MA plot that visualises the difference in gene expression between the GSE37460 and GSE50469 IgAN samples. Assuming that these two set of IgAN samples are essentially represent the same condition, the genes that are differentially expressed in this comparison are the ones that are the most affected by the platform-specific batch effect. Next, we started to remove the genes that are most affected by the batch effects. Sup. Figure 1C visualises Principal component 2 from Supp. Figure 1/A as a function of the gradual removal of the most affected genes, that are represented by the -log10 adjusted p-value of a particular removed affected gene. One can observe a cumulative shrinkage of variance explained by the Principal component 2 due to removal of batch-effect affected genes. Sup. Figure 1D depicts a Principal component analysis (PCA) of gene expression measurements corresponding to the IgAN samples from the two distinct studies post batch effect mitigation. The samples are colored according to their respective study origin. As a result of the batch effect mitigation the principal component 2 shrunk by 4% and the measurements between the two studies seem to be closer to each other than before the procedure. Sup. Figure 1E shows for each gene the variance that is explained by group (CKD entity), study and platform, respectively, after batch effect mitigation. Most genes' variance is explained by group (CKD entity), however, there are genes for which the the variance in expression levels is largely attributed to platform or study. Even though this procedure does not completely eliminate or correct for the batch effects, we could show ,through this example, that we could mitigate the batch effects and gained some confidence that most genes' variance in expression between the samples are due to difference in disease.

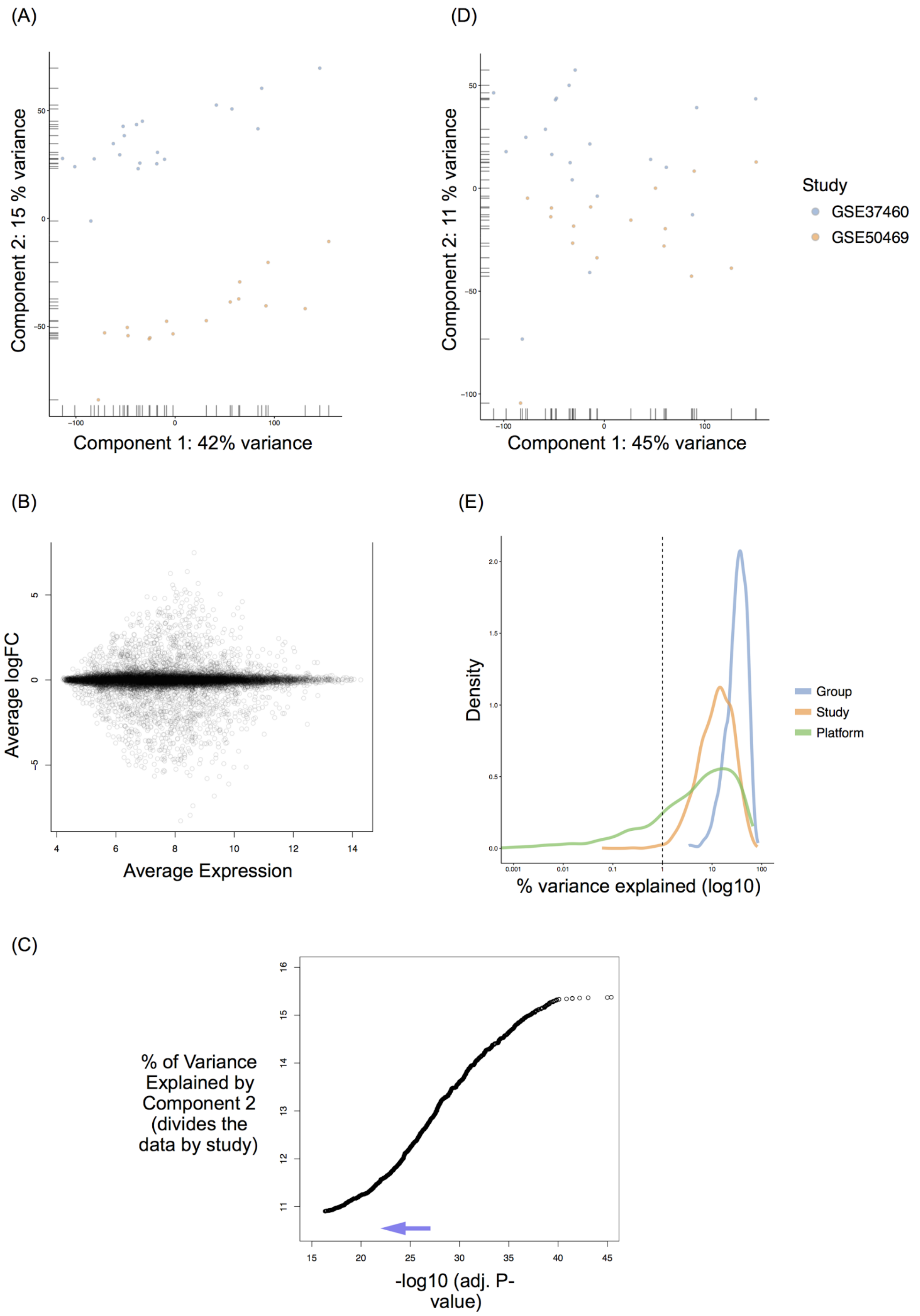
Supplementary Materials

Supplementary Files

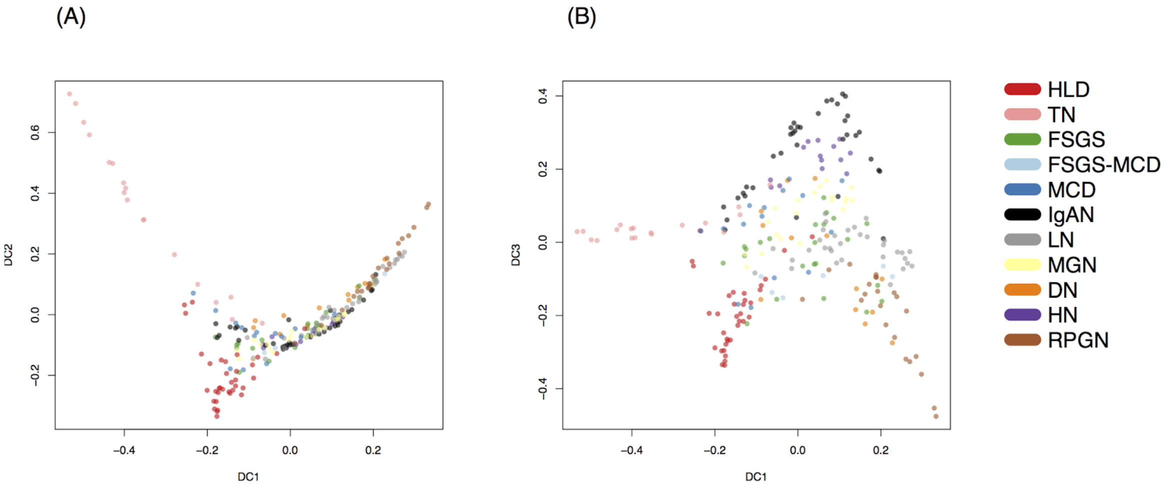
Supplementary file 1.

1790 genes identified by the maxP method [62,63] (see section 4.5.)

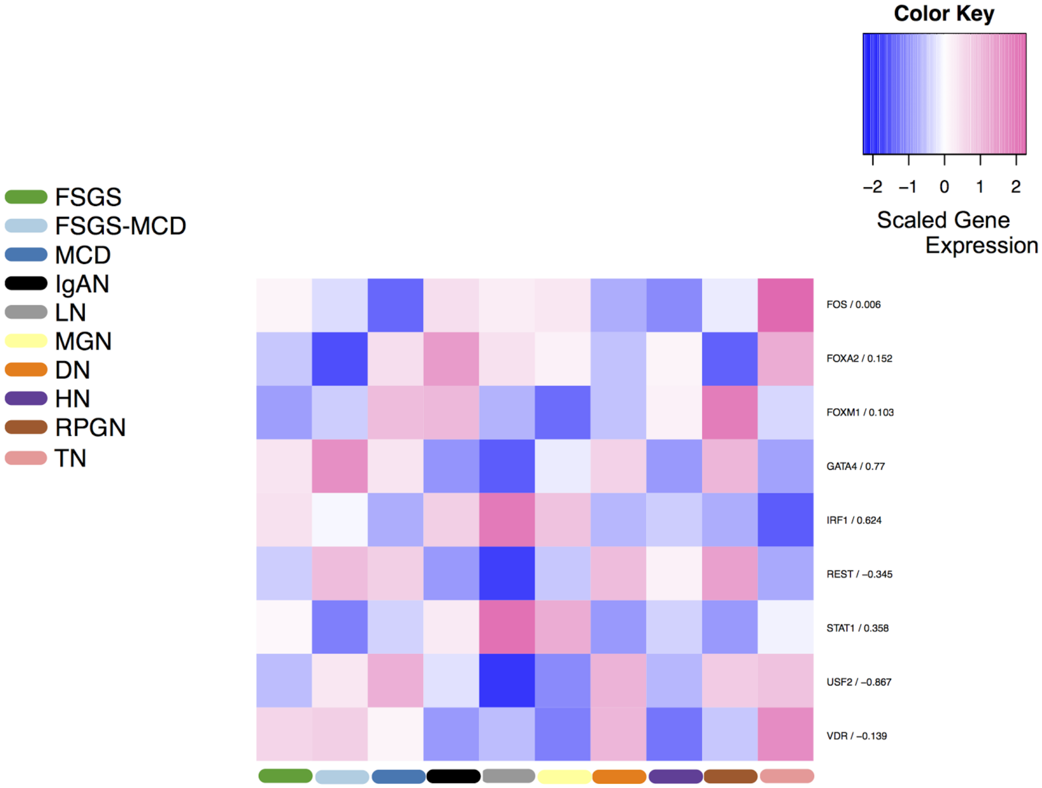
https://drive.google.com/drive/folders/16DUyIfXpDuDjIIYjYOWGHimB2dZirYdg



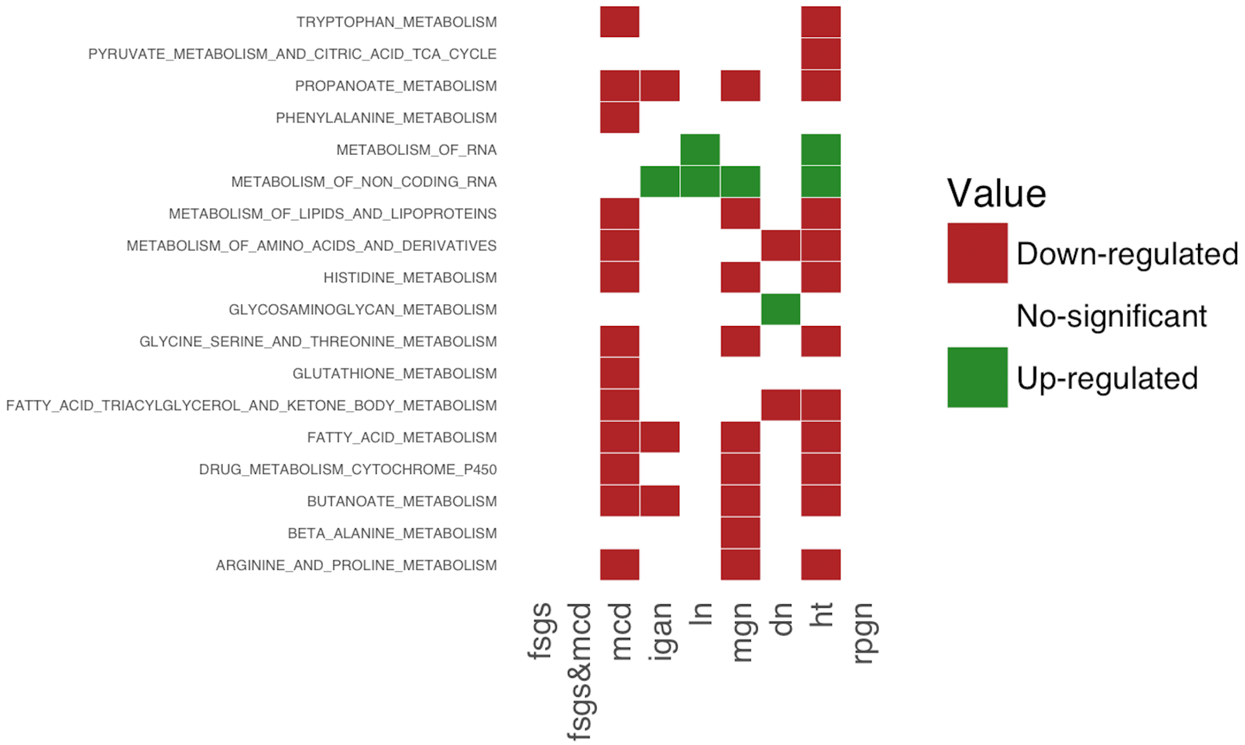
Supplementary Figure 1 - Batch effect mitigation procedure. *(A) Principal component analysis (PCA)* *of gene expression measurements corresponding of IgAN samples from the two distinct studies prior batch effect mitigation. (B) MA plot visualising the difference in gene expression between the GSE37460 and GSE50469 IgAN samples. (C) Principal component 2 (PC2) as a function of the gradual removal of the most affected genes (-log10 adjusted p-value of a particular removed affected gene). (D) PCA of gene expression corresponding to the IgAN samples from the two distinct studies post batch effect mitigation. (E) Depiction of variance for each gene, that is explained by group (CKD entity), study and platform after batch effect mitigation.*



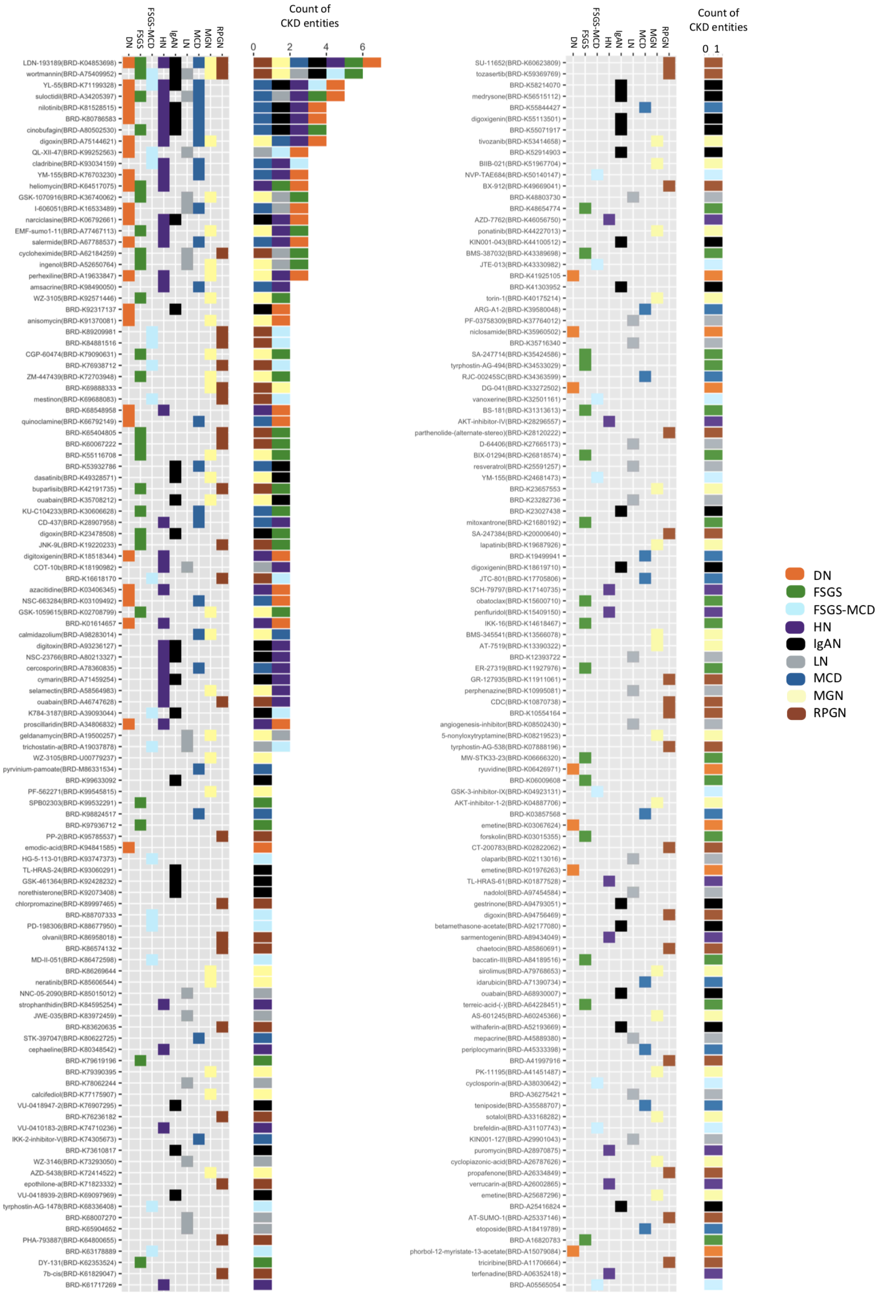
***Supplementary Figure 2 - Two dimensional diffusion maps of CKD entities unravel the geometric trajectory of CKD entities based on their comparative transcriptome profile.** (A) Dimension component 1 (DC1) is depicted against dimension component 2 (DC2), so that the divergence between the controls and the CKD entities are apparent. (B) DC1 is visualised against dimension component 3 (DC3), revealing the fine distinctions between CKD entities.*



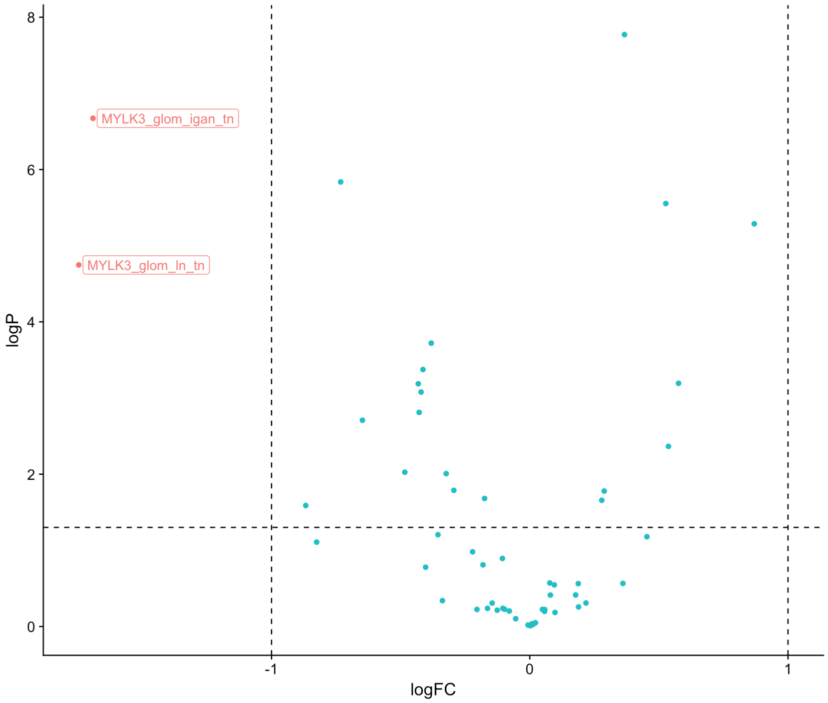
Supplementary Figure 3. *Heatmap depicting the expression of the genes encoding for the transcription factors shown in Figure 4. The expression values were averaged within each condition, then scaled and centered across the conditions. The numbers to the right of factor names are Spearman's rank-based correlation coefficients of factor activity and factor expression across different CKD entities.*



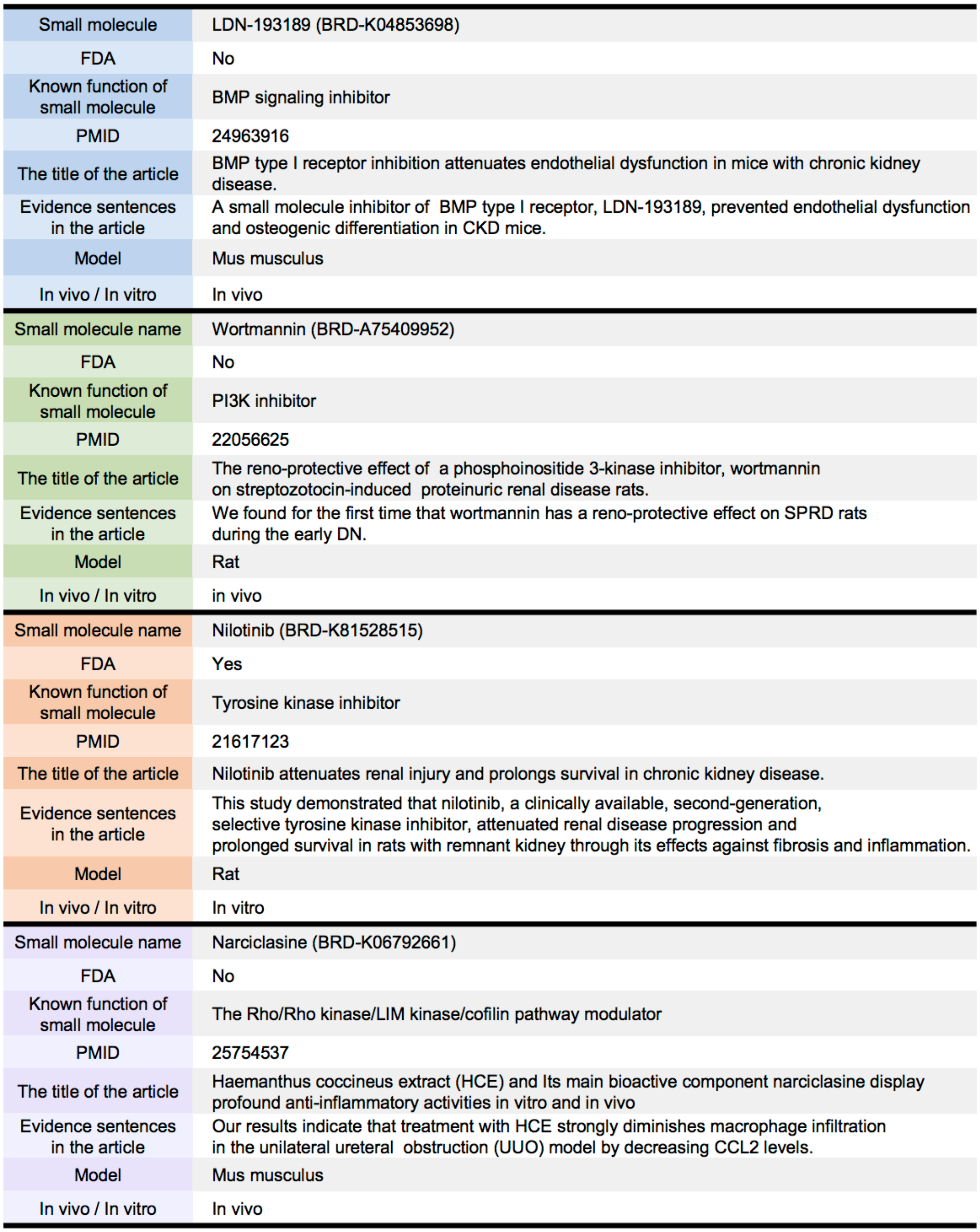
Supplementary Figure 4 - Enrichment of metabolic pathways after gene set analysis. *Pathway analysis result in metabolic pathways ('METABOL'): and their corresponding enrichment: up-regulation (green), down-regulation (red) and non-significant (white). Metabolic pathways are listed in Y axes and disease entities in X axes. Only pathways enriched in at least one disease are shown. Note that FSGS, FSGS-MCD, and RPGN do not have any metabolic pathway significantly affected.*



Supplementary Figure 5. *Bar graph (count of CKD entities) and heatmap of the distribution of 220 small molecules reversely correlated with nine CKD entities. Colored bars on both the bar graph and heat map correspond to the subtype of CKD entities and 220 small molecules are represented on the x-axis of both graphs.*



Supplementary Figure 6. *Volcano plot of differential expression of CKD entities vs TN for glomerular samples for the drug targeted genes. X-axis indicates the log2 of the fold change (FC) and the Y-axis the -log10 of the p-value after differential expression analysis using limma.*



Small molecule	LDN-193189 (BRD-K04853698)
FDA	No
Known function of small molecule	BMP signaling inhibitor
PMID	24963916
The title of the article	BMP type I receptor inhibition attenuates endothelial dysfunction in mice with chronic kidney disease.
Evidence sentences in the article	A small molecule inhibitor of BMP type I receptor, LDN-193189, prevented endothelial dysfunction and osteogenic differentiation in CKD mice.
Model	Mus musculus
In vivo / In vitro	In vivo
Small molecule name	Wortmannin (BRD-A75409952)
FDA	No
Known function of small molecule	PI3K inhibitor
PMID	22056625
The title of the article	The reno-protective effect of a phosphoinositide 3-kinase inhibitor, wortmannin on streptozotocin-induced proteinuric renal disease rats.
Evidence sentences in the article	We found for the first time that wortmannin has a reno-protective effect on SPRD rats during the early DN.
Model	Rat
In vivo / In vitro	in vivo
Small molecule name	Nilotinib (BRD-K81528515)
FDA	Yes
Known function of small molecule	Tyrosine kinase inhibitor
PMID	21617123
The title of the article	Nilotinib attenuates renal injury and prolongs survival in chronic kidney disease.
Evidence sentences in the article	This study demonstrated that nilotinib, a clinically available, second-generation, selective tyrosine kinase inhibitor, attenuated renal disease progression and prolonged survival in rats with remnant kidney through its effects against fibrosis and inflammation.
Model	Rat
In vivo / In vitro	In vitro
Small molecule name	Narciclasine (BRD-K06792661)
FDA	No
Known function of small molecule	The Rho/Rho kinase/LIM kinase/cofilin pathway modulator
PMID	25754537
The title of the article	Haemanthus coccineus extract (HCE) and Its main bioactive component narciclasine display profound anti-inflammatory activities in vitro and in vivo
Evidence sentences in the article	Our results indicate that treatment with HCE strongly diminishes macrophage infiltration in the unilateral ureteral obstruction (UUO) model by decreasing CCL2 levels.
Model	Mus musculus
In vivo / In vitro	In vivo

Supplementary table 1. *Manual curation for four small molecules. For four small molecules, table includes drug name corresponding to four small molecules, biological function, FDA approval status and several informations of articles explaining clinical relevance of small molecules for CKD.*