Molecular disease presentation in diabetic nephropathy

Andreas Heinzel1, Irmgard Mühlberger1, Gil Stelzer2, Doron Lancet2, Rainer Oberbauer3, Maria Martin4 and Paul Perco1

1emergentec biodevelopment GmbH, Vienna, Austria, 2Weizmann Institute of Science, Rehovot, Israel, 3Medical University of Vienna, Vienna, Austria and 4EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, UK

Correspondence and offprint requests to: Paul Perco; E-mail: paul.perco@emergentec.com

ABSTRACT

Diabetic nephropathy, as the most prevalent chronic disease of the kidney, has also become the primary cause of end-stage renal disease with the incidence of kidney disease in type 2 diabetics continuously rising. As with most chronic diseases, the pathophysiology is multifactorial with a number of deregulated molecular processes contributing to disease manifestation and progression. Current therapy mainly involves interfering in the renin–angiotensin–aldosterone system using angiotensin-converting enzyme inhibitors or angiotensin-receptor blockers. Better understanding of molecular processes deregulated in the early stages and progression of disease hold the key for development of novel therapeutics addressing this complex disease. With the advent of high-throughput omics technologies, researchers set out to systematically study the disease on a molecular level. Results of the first omics studies were mainly focused on reporting the highest deregulated molecules between diseased and healthy subjects with recent attempts to integrate findings of multiple studies on the level of molecular pathways and processes. In this review, we will outline key omics studies on the genome, transcriptome, proteome and metabolome level in the context of DN. We will also provide concepts on how to integrate findings of these individual studies (i) on the level of functional processes using the gene-ontology vocabulary, (ii) on the level of molecular pathways and (iii) on the level of phenotype molecular models constructed based on protein–protein interaction data.

Keywords: biological pathways, diabetic nephropathy, functional annotation, molecular disease modeling, omics profiling

INTRODUCTION

Kidney disease is a common problem in patients with diabetes. Depending on the degree of GFR impairment and/or albuminuria, roughly 30% of subjects are affected. As with most diabetes-associated comorbidities, the pathophysiology is multifactorial and the molecular pathways involved in the initiation and progression constitute a wide and complex as well as redundant network of regulators. Considerable success has been achieved in uncovering certain aspects of molecular regulation of these processes. The most prominent examples are the identification of the causal contribution of the renin–angiotensin–aldosterone system (RAAS) as well as the matrix metalloproteinase network (MMP) [1, 2]. However, this gain in knowledge did not lead to an equal success in the discovery of actual drugs that could prevent or slow down the pace of this progressive disease. The only exception is the invention of the blockade of the RAAS at several different sites by using angiotensin-converting enzyme (ACE) inhibitors or angiotensin-receptor blockers (ARBs) [3–5].

There are several valid explanations for this uncomfortable situation but probably the most striking cause lies in the multifactorial and highly redundant nature of the molecular pathways of this disease. The revolutions in information technology as well as the omics technologies, however, offer nowadays new opportunities to uncover these complex molecular signal cascades thus laying the ground for potentially uncovering diagnostic and prognostic as well as therapeutic leads to tackle this highly prevalent clinical problem. This review highlights the most recent research in this area and provides rational strategies of how to proceed in the fight against diabetic nephropathy (DN). If at
least some of these in-silico concepts of rational marker and drug design will prove successful in clinical trials then a major step forward will have been made towards an improved care of patients with diabetes. Needless to say, such concepts have the potential to be applied to other multifactorial entities and are by no means limited to DN. In this manuscript, we will specifically outline key omics studies conducted in the area of DN as well as analyse concepts on a functional level, on the level of molecular pathways and on the level of biological networks being applied to get a deeper understanding of the pathophysiology of DN.

OMICS studies in diabetic nephropathy

A number of omics studies have been conducted in the past couple of years with a focus on DN, as also indicated by the MEDLINE query ‘diabetic nephropathies[majr] AND (gene expression profiling[mh] OR microarray analysis OR proteomics[mh] OR metabolomics[mh] OR genome-wide association study[mh]) NOT review’ resulting in 151 publications. Representative studies with a focus on disease pathophysiology for each omics category were selected for this review, also complemented by the authors’ personal paper collections as well as selected references of key articles. Supplementary Data, Table 1 holds the discussed omics studies with information on biological processes presented in the respective Results and Discussion sections. The reader is also referred to the succeeding articles within this special issue dedicated to the individual omics tracks for an extended discussion of performed studies.

Genomics studies in DN. A significant number of Genome-Wide Association Studies (GWAS) in the context of DN has concentrated on patients of African ancestry [6, 7], who have a disproportionately high risk for developing type II diabetes mellitus [8]. Single-nucleotide polymorphisms in the MHY9 gene are known to be highly associated with ESRD in this population [9, 10] and could also be linked to DN [6, 7]. More recently, it was shown that MHY9 variants also mediate DN susceptibility in European Americans [11]. Other SNPs discussed in the context of DN are located in the FMRD3 gene [6, 12], and results from GWAS based on samples from the multi-ethnic Family Investigation of Nephropathy and Diabetes (FINDD) study include SNPs near the genes CNDP1 [13], CACNB2, ARL5 and NEBL [14].

In search of genetic determinants explaining the lower incidence of DN in diabetic women, a GWAS in the Finnish Diabetic Nephropathy (FinnDiane) cohort identified variants in a region between the two transcription factors SP3 (which is known to interact with the estrogen receptor alpha) and CDCA7 [15]. The reader is referred to the manuscript on ‘genome-wide studies to identify risk factors for kidney disease with a focus on patients with diabetes’ by Regele et al. within this special NDT issue for a comprehensive overview on identified SNPs in DN.

Transcriptomics studies in DN. More than a decade ago, the results of the first cDNA microarray study on DN, in a set of 81 genes differentially expressed in the kidneys of healthy and (STZ)-induced diabetic mice, were published by Wada et al. [16]. Although these results contributed to a molecular understanding of the role of hyperglycaemia in the early progression of glomerulosclerosis, a significant drawback was the lack of annotation for more than a half of the identified molecules at that time.

A few years later, the first oligonucleotide arrays were available to study transcriptional changes with an increased coverage of the full transcriptome, and new players as, for example, members of the Wnt signalling pathway, were identified as contributors to the pathophysiology of DN [17].

A recurring conclusion in published DN transcriptomics studies is the importance of oxidative stress underlying fibrosis for DN progression. Morrison et al. [18] found an up-regulation of several thiol antioxidative genes in rat mesangial cells treated with high glucose concentrations, and Cheng et al. [19] identified increased concentrations of members of the thioredoxin system under comparable experimental conditions in mice.

While in-vitro and animal studies mainly concentrated on the glomeruli, transcriptomics studies of microdissected human kidney biopsies from DN patients showed relevant changes in gene-expression profiles also in the tubulointerstitial compartment, including several nuclear factor-kappaB targets as well as genes linked to inflammation [20, 21] and angiogenesis [17]. Comparisons of compartment-specific gene-expression changes identified common deregulated mechanisms in glomeruli and tubuli associated with cell proliferation [22] and Janus kinase signalling [23], pointing towards the role of hypertrophy and fibrosis in DN.

Meanwhile, transcriptomics studies on DN are no longer confined to the protein-coding regions of the genome. Argypoulous et al. [24] used miRNA arrays and could identify urinary miRNA profiles that differ across the different stages of DN.

Proteomics studies in DN. The majority of proteomics studies in DN focus on the identification of biomarker panels. In particular, studies in the urine of DN patients resulted in a number of potential markers showing altered concentration when compared with healthy controls like, e.g. the ribosomal ubiquitin fusion protein UBA52 [25] or alpha 1 antitrypsin, the latter also showing an up-regulation in diseased kidney biopsies in areas of renal fibrosis [26].

Jin et al. [27] reported on a diagnostic panel of three potential markers, namely alpha-1-antitrypsin, alpha-1-acid glycoprotein 1 and prostate stem cell antigen. Zuerbig et al. [29] showed that the previously identified prognostic classifier for chronic kidney disease consisting of 273 urinary peptides (CKD273) [28] was capable of predicting development of nephropathy in a diabetes cohort before the onset of microalbuminuria.

So far, proteomics studies on kidney tissue in the context of DN are rare and mainly conducted in animal models. Identified molecules are, e.g. PPARalpha which was identified as a hub protein in the set of differentially expressed proteins in diabetic animals [30] as well as a number of nephropathy-specific proteins including the phospholipids scramblase 3 and tropomyosin 3 in the whole kidney tissue of STZ rats [31].

Metabolomics studies in DN. Since metabolomics is one of the latest advances in the field of omics, not least because of the technical challenges regarding the chemical diversity and the
wide range of concentrations of metabolites, the number of studies using non-targeted approaches for metabolomics signature detection in the context of DN is small with blood (either serum or plasma) being the sample matrix of choice in most studies.

Identified molecular processes include amino acid and phospholipid metabolism [32] or purine metabolism [33]. Han et al. [34] found an association in concentration levels of esterified and non-esterified fatty acids in plasma and different stages of DN.

Results from urine metabolomics studies in DN patients could elucidate different disease aspects than serum or plasma studies, e.g. the role of mitochondrial metabolism [35], specific aspects of altered tryptophan metabolism that were detected in a subset of the FinnDiane cohort [36] or the contrary tendency of citric acid cycle intermediates in urine from db/db mice when compared with serum profiles [37].

Hirayama et al. [38] identified a signature of five metabolites in serum capable of discriminating diabetic patients with and without macroalbuminuria, including γ-butyrobetaine, symmetric dimethylarginine, azelaic acid and two unknown metabolites.

Methods and tools for integrating findings from individual studies will be discussed in the next chapters contributing to a better understanding of molecular pathophysiology in DN, mainly focusing on functional annotation of molecular features, identification of deregulated molecular pathways as well as identifying deregulated molecular processes on the level of constructed biological networks (Figure 1).

Gene ontology and the Renal Gene Ontology Annotation Initiative

The Renal Gene Ontology Annotation (Renal GOA) Initiative [http://www.ebi.ac.uk/GOA/kidney] was established as an effort to summarize the accumulated experimentally based information in a comprehensive public resource of high-quality functional annotation for proteins involved in renal development, function and disease [39]. These annotations were based in the structured Gene Ontology (GO) [http://geneontology.org/] vocabulary for both (i) improving the descriptiveness of terms representing renal processes and (ii) increasing the number of associations of proteins involved in the renal system to informative GO terms. The initiative was a 3-year project funded by Kidney Research UK and based in the European Bioinformatics Institute (EMBL-EBI), aiming to provide a unique public resource of comprehensive high-quality functional annotation for proteins involved in renal development, function and disease.

GO uses structured controlled vocabulary terms to describe three aspects of a gene product: the ‘molecular function(s)’, or the activity(ies) that the sequence can directly perform; the ‘biological process(es)’ or ‘the operation(s) it’ contributes to; and the subcellular locations (‘cellular components’) in which it is located. GO terms are organized into a graph, where each term is linked to one or more general ‘parent’ terms and one or more specific ‘child’ terms if applicable. For example, the GO term ‘nephron morphogenesis’ (GO:0072028) has two parent terms, ‘kidney morphogenesis’ (GO:0060993) and ‘nephron development’ (GO:0072006), and 12 child terms, including ‘renal vesicle morphogenesis’ (GO:0072077) and ‘mesonephric nephron morphogenesis’ (GO:0061228) (Figure 2).

The GO annotation data set provided by the GO Consortium is one of the most widely used resources in biomedical and biotechnological data analysis, assisting researchers in interpreting, validating and forming hypotheses for their data [40]. The Renal GOA initiative allowed (i) generation of detailed manual GO annotation for renal systems and (ii) to develop and improve the terms in the Gene Ontology to ensure that the whole of renal biology is well represented. The data set is available through GO and provides consistent annotations for renal-specific data sets enabling researchers to rapidly evaluate and interpret existing renal data and generate hypotheses to guide future research with confidence. This initiative has associated 163 156 GO terms to 2810 distinct UniProtKB proteins from the prioritized renal-related list. Of these, 1025 prioritized proteins have been annotated using GO terms. Additionally, over 600 new renal-specific GO terms have been created.

We annotated the DN omics studies listed in the previous chapters (Supplementary Data, Table 1) with GO biological process terms reflecting the major outcomes of the particular work. Specific terms were assigned whenever (i) the process itself or (ii) key molecular features of a process were the main subject of the Results or Discussion section in the respective article. Results in the context of ‘extracellular matrix organization’ were discussed in 14 omics studies followed by the GO term ‘immune response’ which was discussed in 10 studies. Among the top-ranked terms based on paper counts were also ‘response to oxidative stress’, ‘angiogenesis’ or ‘apoptosis’. A graphical representation within the GO hierarchy of those GO terms, which were discussed in at least three of the omics studies, is depicted in Figure 3A.

Pathway representation and superpathways

Next to studying omics result profiles on the level of GO terms, protein to molecular pathway assignments as stored in molecular pathway databases can be used in order to identify deregulated mechanisms on a functional level. Prominent pathway databases include KEGG [http://www.kegg.jp], Reactome [http://www.reactome.org], or Wiki pathways [http://www.wikipedia.org] with recent efforts trying to consolidate pathways from various sources, such as PathwayCommons [http://www.pathwaycommons.org/], NCBI’s Biosystems [http://www.ncbi.nlm.nih.gov/biosystems/], Human Pathway Database (HPD) [http://discoveryinformatics.iupui.edu/HPD/], ConsensusPathDB [http://consensuspathdb.org/] or PathJam [http://www.pathjam.org/]. The GeneCards system currently holds information on molecular pathways from 12 sources covering a total of 3215 biological pathways and 11 478 unique genes [41]. The four largest pathway sources (KEGG, Reactome, Wiki pathways and Qiagen) used in GeneCards collectively contain 10 770 genes with 1413 being shared between all four sources.

Within GeneCards, the main pathway clustering algorithm is based on the Jaccard similarity coefficient (J), which scores the degree of gene sharing for two pathways taking into account their combined pathway size. In the clustering approach, two separate algorithms are combined. One is keeping edges
between each pathway and its best matching pathway (i.e. nearest neighbour), thus allowing grouping together two pathways not necessarily having a high Jaccard score but are the best matching (this is specifically important to connect similar pathways from different databases where the gene sets used could be very different and thus lower the Jaccard score). The second algorithm implemented in the approach is a hierarchical clustering algorithm with a relatively high threshold (this is important to connect highly similar sets of genes that mainly originate from the same database where slightly different sets of genes are annotated as a separate pathway). The reason for using a combination of these algorithms is in order to, on the one hand, reduce the number of singletons which are determined by the threshold used in the nearest neighbour algorithm (e.g. when the highest score of a pathway is 0.3), yet still keep highly similar pathways in the same cluster when the best matching pathway is very similar (e.g. if two pathways have a Jaccard score of 0.99 and there is another pathway with a Jaccards score of 0.95). This approach resulted in a set of 1073 unified pathways, termed SuperPaths, that connect maximally informative pathway-related annotations to every human gene [42]. By judiciously supervising the resulting SuperPath size (i.e. the number of its contained pathways), a high measure of annotation specificity was preserved while minimizing redundancy.

SuperPaths are available in the GeneCards pathway section enabling a clear depiction of the set of unified pathways for each gene [41]. A new GeneCards companion database, called PathCards [http://pathcards.genecards.org/], was constructed to

FIGURE 1: Schematic analysis workflow for delineating molecular DN disease presentations. With the phenotype at the centre, data are gathered from the scientific literature as well as from different omics studies in order to generate a phenotype molecular feature set. Bioinformatics analyses on the level of GO term enrichment analysis, molecular pathways as well as via identification of process units leads to the identification of deregulated molecular processes and pathways as well as the construction of molecular models for the given disease.
present a web card for quick in-depth analysis of each human SuperPath in addition to basic search capabilities in a pathway-centric orientation. This online database enables a pathway connectivity view for each SuperPath as well as access to the gene lists of the SuperPath and of each of its constituent pathways.

Metabolic pathways are composed of genes whose proteins are mostly connected via compounds that serve as substrates or products, whereas signalling pathways are characterized by proteins that directly interact and thereby transfer signals within cells. This prominent difference can clearly be seen using the interaction network derived from STRING that is displayed in PathCards. Upon elucidation of DN-associated SuperPaths, compounds found in metabolic pathways might be used as remedy targets, whereas signalling pathways might elicit a treatment approach for altered protein–protein interactions (PPIs). The list of top-ranked SuperPaths based on a set of human omics-derived molecular features showing the highest fold-changes in the context of DN are, e.g. ‘Cell adhesion_ECM remodeling’ or ‘Degradation of the extracellular matrix’ [43]. In addition, various signalling SuperPaths were associated with deregulated molecular features such as ‘VEGF signalling pathway’, ‘PEDF Induced Signalling’ and ‘Chemokine Signalling’, thereby indicating the complex interplay of signals involved in processes leading to DN.

We previously extracted molecular pathways from scientific articles focusing on DN [44]. A pathway map was constructed based on the set of 27 identified molecular KEGG and PANTHER pathways where edges between individual pathways were based on shared molecular features (Figure 3B).

**Biological networks and molecular models of disease**

Next to evaluating proteins in the context of molecular pathways as listed in the previous section, researchers started to use interaction data sets such as PPIs in order to delineate relations...
on a molecular level thus being able to expand analyses to proteins currently not assigned to a dedicated pathway but being linked to other molecules via an interaction data.

A PPI depicts a physical interaction of different strength and evidence between two proteins. Such interactions do not necessarily imply that there is any functional relationship between
the partners, but merely indicate that two proteins bind to each other (e.g. direct physical interaction) or at least are found in close proximity to each other (e.g. co-localization). The establishment of technologies such as co-immunoprecipitation, yeast two-hybrid screens or affinity electrophoresis facilitating high-throughput screening of PPIs made interaction data for model organism as well as human readily available. These data can be obtained from publicly accessible databases such as IntAct [http://www.ebi.ac.uk/intact/] or BioGRID [http://thebiogrid.org/].

Next to PPIs, there are other types of networks such as (i) transcriptional regulatory networks specifically focusing on the regulation of DNA target gene transcription through transcription factors, (ii) 'genetic interaction networks', which are composed from observations on simultaneous alteration of two (or more) genes, e.g. by mutations causing an unexpected change in the resulting phenotype given the effect of each individual alteration alone, (iii) 'metabolic networks', which aim at capturing the complete set of metabolic reactions in an organism by recording the interplay between enzymes, transporters and metabolites and their involvement in reactions transforming substances and metabolites into specific substances, (iv) 'co-expression networks', which focus on relations between genes showing similar expression patterns over a set of samples, or (v) 'co-annotation networks', which focus on co-mentioning of two proteins in the scientific literature. For an extensive review on different types of biological networks, the reader is referred to Vidal et al. [45].

Next to network visualization and analysis tools such as Cytoscape [http://cytoscape.org/], Osprey [http://biodata.mshri.on.ca/osprey/], of VisANT [http://visant.bu.edu/], algorithms and methodologies have been developed in order to expand the set of experimentally derived PPIs based on predictions taking into account information on pathway assignment, structural protein domain information or functional protein annotation among others [46, 47]. The STRING database integrates experimentally determined and predicted protein interactions to provide a comprehensive compendium of direct as well as indirect protein interactions for more than 1000 organisms [http://string-db.org/].

Biological network analysis has also been applied in the context of DN as outlined in the examples given below. Using visual network inspection methods as well as basic quantitative measures, Kretzler et al. [48] could show in a network build from expression data on seven different chronic renal disease entities that the current clinical disease classification does not reflect molecular similarities on a molecular level. Fehete et al. [49] demonstrated that while only minor direct feature overlap between molecular features implicated in the pathophysiology of DN and those reported to be of clinical relevance existed, a molecular feature landscape generated by consolidating those features on an expanded pathway network reflected the clinical implications of DN. Agrawal et al. [50] used sub-networks identified in a network specifically constructed to cover type 2 diabetes and related complications by combining PPIs with co-expression information to elaborate on the development of diabetes and associated complications such as DN. Grekas et al. [51] identified functional network modules as well as hub and bottleneck genes that due to their nature are likely to play central roles in a PPI network of DN-associated genes. He et al. [52] were interested in glomerular-specific molecular processes and constructed a PPI network in the glomerulus which they termed GlomNet in order to find relevant signalling networks and molecules of relevance in glomerular development and disease. Tang et al. [22] recently constructed a gene co-expression network based on transcriptomics data in order to identify key regulators of transcription both in tubuli and glomeruli of diabetic nephropathy samples when compared with samples from control subjects. Within the SysKid project, we constructed a molecular process model of DN using the omicsNET dependency network as underlying biological network for multi-omics data integration, hypothesis generation, and biomarker selection [53]. In brief, molecular features reported as being differentially regulated in (i) genomics, (ii) transcriptomics, (iii) proteomics and (iv) metabolomics experiments on DN were retrieved. The set of omics-derived molecular features was complemented by a set of genes extracted via an automatic literature search for diabetic nephropathy-associated genes. The set of in total 2466 features was mapped on the aforementioned omicsNET biological network constructing a disease network by including direct edges between features of the input set and discarding molecular features not connected to any other features of the input list. The resulting DN-specific networks held 1921 molecular features with 28 748 interactions and served as input to the MCODE algorithm for identifying clusters of tightly interconnected molecular features following the model forming procedure as outlined in Heinzel et al. [54]. Thirty-four MCODE clusters were identified holding in total 688 proteins (Figure 3C). The network molecular model was used in order to delineate biomarker candidates for the most relevant process units. The set of selected biomarkers was shown to improve prediction of accelerated eGFR decline with a median follow-up time of 4 years in patients with type 2 diabetes on top of established clinical risk factors [55]. A validation study in an independent cohort of about 2000 patients with type 2 diabetes to assess performance of the marker panel is currently ongoing.

Comparison of molecular models via feature overlap of key network components depicts another way of analysing biological networks like in the study by Heinzel et al. [56] on the analysis of common mechanisms between chronic kidney disease and cardiovascular disease.

Biological network interference analysis is also a way to delineate on-target as well as off-target drug effects as well as supporting the identification of predictive biomarkers on drug response as exemplified by Heinzel et al. [44] in a recent study where the analysis of model overlap between a molecular model of DN and a drug mechanism of action molecular model for the group of ACE inhibitors was primary focus of the study. A set of seven biomarkers is in the end proposed serving as proxies for the interference of the ACE-inhibitor mechanism of action molecular model and the molecular model for diabetic nephropathy.

Outlook and perspectives
The wealth of publicly available omics data sets has opened new avenues of studying pathophysiological mechanisms in
renal diseases and especially DN as the most prevalent form of chronic renal diseases. Whereas the first omics studies focused on the identification of the most differentially regulated individual molecular entities, researchers have in the meantime adopted this concept of integrated analysis in order to (i) identify pathological processes in disease development and progression, (ii) derive novel biomarker candidates and drug targets as well as (iii) study drug on-target and off-target effects. The need for a better prediction of drug effects on a molecular level was recently outlined by Heerspink et al. [57] in order to increase success regarding drug efficacy in clinical trials but also to stop development of compounds which trigger adverse side effects down the road by interfering with molecules in the molecular network off the prime target. We are currently just at the beginning of making full use of the available data and with improved methodologies for analysis and additional omics profiles on clinically well-characterized samples, success rates of clinical trials in drug development will hopefully again increase in the near future.

**SUPPLEMENTARY DATA**

Supplementary data are available online at http://ndt.oxfordjournals.org.

**ACKNOWLEDGEMENTS**

The research leading to these results has received funding from the European Community’s Seventh Framework Programme under grant agreement no. HEALTH-F2-2009-241544 (SysKid).

**CONFLICT OF INTEREST STATEMENT**

Results presented in this paper have not been published previously in whole or part, except in abstract format. A.H., I.M. and P.P. are employees of emergentec biodivelopment GmbH.

**REFERENCES**


Received for publication: 12.1.2015; Accepted in revised form: 6.6.2015