Mapping of molecular pathways, biomarkers and drug targets for diabetic nephropathy

Raul Fechete[1], Andreas Heinzel[1], Paul Perco[1], Konrad Mönks[2], Johannes Söllner[1], Gil Stelzer[3], Susanne Eder[2], Doron Lancet[3], Rainer Oberbauer[4], Gert Mayer[2] and Bernd Mayer[1,5]

[1]emergentec biodevelopment GmbH, Vienna, Austria

[2]Department of Internal Medicine IV (Nephrology and Hypertension), Medical University of Innsbruck, Innsbruck, Austria

[3]Department of Molecular Genetics, Weizmann Institute of Science, Rehovot, Israel

[4]KH der Elisabethinen Linz and Medical University of Vienna, Vienna, Austria

[5]Institute for Theoretical Chemistry, University of Vienna, Vienna, Austria

Correspondence:

Dr. Bernd Mayer, emergentec biodevelopment GmbH, Gersthofer Strasse 29-31, 1180 Vienna, Austria

E-mail: bernd.mayer@emergentec.com

Fax: +43 1 4034966-19

Abbreviations:

ACEI: angiotensin-converting enzyme inhibitor

ARB: angiotensin II type 1 receptor blocker

DN: diabetic nephropathy

GO: Gene Ontology

GWAS: Genome Wide Association Study

GSEA: Gene Set Enrichment Analysis

KEGG: Kyoto Encyclopedia of Genes and Genomes

MeSH: Medical Subject Heading

MIA: microalbuminuria

NIH: National Institute of Health

PPI: protein-protein-interaction

SNP: single nucleotide polymorphism

Keywords:

Biomarker / Data integration / Modeling / Network / Target

Total number of words:

(including references, table captions, figure captions): 9,399

Clinical Relevance

Technical maturity and procedural standardization of Omics, prominently including transcriptomics, proteomics and metabolomics, has introduced these tools in the clinical context aimed at patient specific disease profiling. As a next step, integration of heterogeneous Omics profiles across the molecular domains promises identifying relevant processes and pathways not becoming evident on a single Omics level. This approach appears of particular relevance for complex diseases such as diabetic nephropathy demanding a representation of chronic kidney disease in the context of diabetes and hypertension.

Multilevel data consolidation and integration for diabetic nephropathy, involving next to Omics profiles also literature, patent and clinical trial text mining provides a heterogeneous spectrum of molecular features. This alleged heterogeneity vanishes when interpreting the features in the context of pathways and protein interaction networks, where distinct sets of disease associated processes become evident. Identification of such disease associated pathways furthermore allows to link pathway-specific biomarkers and drug targets. This integration concept proves valid in representing the intricate interplay of diabetic nephropathy, cardiovascular disease and diabetes on a molecular network level, in turn offering a platform for analysis of biomarkers and drugs in the context of specific Omics profiles characterizing diabetic nephropathy.

ABSTRACT

Purpose:

For diseases with complex phenotype such as diabetic nephropathy integration of multiple Omics sources promises an improved description of the disease pathophysiology, being the basis for novel diagnostics and therapy, but equally important personalization aspects.

Experimental design:

Molecular features on diabetic nephropathy were retrieved from public domain Omics studies and by mining scientific literature, patent text and clinical trial specifications. Molecular feature sets were consolidated on a human protein interaction network, and interpreted on the level of molecular pathways in the light of the pathophysiology of the disease and its clinical context defined as associated biomarkers and drug targets.

Results:

About 1,000 gene symbols each could be assigned to the pathophysiological description of diabetic nephropathy and to the clinical context. Direct feature comparison showed minor overlap, whereas on the level of molecular pathways the complement and coagulation cascade, PPAR signaling, and the renin-angiotensin system linked the disease descriptor space with biomarkers and targets.

Conclusion and clinical relevance:

Only the combined molecular feature landscapes closely reflect the clinical implications of diabetic nephropathy in the context of hypertension and diabetes. Omics data integration on the level of interaction networks furthermore provides a platform for identification of pathway-specific biomarkers and therapy options.

1 Introduction

Diabetic nephropathy occurs in both, type 1 and 2 diabetes mellitus. In patients with type 1 disease approximately 20 to 30 percent of affected individuals will develop microalbuminuria as the first clinical sign of renal disease after a median diabetes duration of about 15 years [1]. Early detection of these individuals is crucial as it has been shown that less than half of the patients will further progress to overt nephropathy (defined as urinary albumin excretion > 300 mg/day). This finding is mainly due to better metabolic control, more aggressive blood pressure reduction, and the use of agents blocking the renin-angiotensin system [2, 3]. In type 2 diabetes robust data on the incidence of nephropathy were derived from the United Kingdom Prospective Diabetes Study: Among the 5,100 patients with newly diagnosed diabetes the prevalence of microalbuminuria, macroalbuminuria and either an elevated plasma creatinine concentration or requirement of renal replacement therapy after 10 years was 25%, 5% and 0.8%, respectively [4]. However, as the prevalence of type 2 diabetes for all age groups worldwide is expected to rise from 2.8% seen in the year 2000 to 4.4% in 2030, which corresponds to an increase in absolute patient numbers from 171 million to 366 million, it is not surprising that already at present diabetic nephropathy is by far the leading cause of end stage renal disease [5]. Of particular importance is the fact that many patients die because of cardiovascular disease before reaching dialysis [4].

The early diagnosis of diabetic nephropathy rests on the measurement of protein and/or albumin excretion in urine, and microalbuminuria (MIA, excretion of 30-300 mg albumin per day) is currently considered as the diagnostic gold standard. In patients with type I diabetes MIA has an excellent sensitivity as well as specificity to identify patients at risk for the development of more severe nephropathy, and a reduction of urinary albumin excretion is associated with the preservation of the glomerular filtration rate. In patients with type 2 diabetes, which form the majority of subjects in clinical practice, the specificity of MIA for diabetic nephropathy is much lower even though it is an established ominous sign for an adverse cardiovascular prognosis [1, 6]. This lack of specificity however leads to treatment problems. Whereas blockade of the renin-angiotensin system still is very effective in preventing progression of renal disease in patients where MIA is a sign of early nephropathy, the same therapeutic intervention did not preserve glomerular filtration rate in the ONTARGET study, which included subjects where MIA was more an indicator of increased cardiovascular risk [7]. Accordingly, in the ACCOMPLISH trial the combination therapy of benazepril and hydrochlorothiazide reduced albuminuria in hypertensive subjects but resulted in an almost doubled rate of chronic kidney disease compared to benazepril/amlodipine therapy [8]. These data led the participants of a NKF and FDA (National Kidney Foundation and US Food and Drug Administration) sponsored meeting to the conclusion that proteinuria in fact is only a surrogate outcome in kidney disease progression [9]. Multiple risk factors for the development of diabetic nephropathy have been identified. These include amongst others genetic susceptibility, age, race, obesity, smoking, blood pressure, glomerular hypertrophy and hyperfiltration, hyperglycemia, and the formation of advanced glycation end products as well as cytokine activation [10].

Some of these risk factors can be modified by therapy and thus current treatment regimen are directed against systemic and especially intraglomerular hypertension (blood pressure lowering medication and blockade of the renin-angiotensin-aldosterone system), as well as strict glycemic control [11-15]. However, as cardiovascular mortality is also a major focus of treatment a multifactorial approach, which additionally includes smoking cessation, dietary and behavior modification [16], lipid lowering therapy and administration of aspirin is recommended [17, 18].

Numerous Omics technologies have been applied for deciphering biological processes linked to DN, including genome-wide association studies [19], transcriptomics [20], proteomics [21] and metabolomics [22]. Traversing Omics screening results towards identification of biomarkers is seen with proteomics: Rossing and colleagues applied proteomics profiling in a cohort of over 300 patients for studying diabetic nephropathy and non-diabetic kidney diseases [23], and the method was recently expanded in a multicenter validation study for identification of subjects with DN resting on detection and quantification of urinary collagen fragments [24].

Further integrating such data (cross-Omics) leading to multi-marker models to be used for diagnosis or disease monitoring is seen as future perspective, as recently outlined by Fox et al. [25]. More generally, DN may be considered as ideal case for utilizing the concept of Systems Biology [26], as complex pathophysiology is seen for the kidney in the realm of diabetes mellitus and hypertension, in turn closely linking to bone metabolism and cardiovascular implications denoted as the cardiorenal syndrome. Cross-omics integration of Omics data from kidney transcriptomics and proteomics, together with literature annotation of molecular features relevant in the cardiovascular context identified the coagulation pathway as important cardiorenal process [27, 28]. In this work protein-protein interaction networks (PPIs) were used for studying causative as well as associative dependencies of feature profiles and clinical indications. In the last years the analysis of Omics data on gene or protein interaction networks has been established as de-facto standard for multidimensional data mapping and interpretation [29], and concepts for linking molecular data and clinical phenotype space have emerged. Hidalgo et al. [30] introduced the Phenotypic Disease Network (PDN) as a map summarizing phenotypic connections between diseases. Barrenas et al. employed disease-gene networks [31] with specific focus on topological characteristics of such graphs. Shortest path-based algorithms (among others) are frequently applied for linking multilevel networks aimed at identifying key regulatory genes and proteins [32-35], but also for identification of functional modules and pathways affected in the diseased state [36]. Furthermore, concepts for building disease-specific drug-protein connectivity maps have been introduced [37], altogether aimed at building relations between molecular feature space, clinical data space, and associated markers, drug targets and associated drugs.

In the present work we integrated data sources characterizing DN on a molecular level by utilizing a full human proteome interaction network as common denominator, specifically including data coming from DN case-control Omics studies. We further annotated the network with DN-associated data retrieved from scientific literature, patent text and clinical trial references, and then used this annotated PPI for linking the pathophysiological description of DN with biomarker candidates and therapy targets discussed for diagnosis, prognosis and therapy of DN.

2 Materials and methods

2.1 Data sets, Omics studies

A search in PubMed (http://www.ncbi.nlm.nih.gov/pubmed, database status as of June 2010) was performed in order to identify publications utilizing Omics in the context of DN. Next to the clinical phenotype “diabetic nephropathy” the keywords “microarray”, “transcriptomics”, “proteomics”, “metabolomics”, “metabonomics” or “SNP” were applied. This retrieval of Omics studies resulted in 146 publications, of which 31 specifically reported Omics screening results and features deregulated in DN as identified in a case-control study design. Of these, 17 were done with human specimens further considered, comprised of 4 transcriptomics, 7 proteomics, 2 metabolomics and 4 SNP studies, as referenced in Table 1.

Insert Table 1 here

SNPs found in open reading frames of protein coding genes were directly mapped to the respective gene ID, or otherwise assigned to the gene ID of the next transcription start site of a protein coding reading frame, all according to the ENSEMBL human genome representation. Metabolites were mapped to associated enzymes using the Human Metabolome Database (HMDB) [38] (status as of September 2010). Across all four Omics domains 851 unique gene symbols could be retrieved.

2.2 Data sets, literature mining

Features associated with DN were furthermore extracted via literature mining. A PubMed search using the MeSH heading “diabetic nephropathy” was applied in conjunction with the gene2pubmed association file (ftp://ftp.ncbi.nih.gov/gene/DATA/gene2pubmed.gz, status as of July 2010) maintained by NCBI enabling a gene-to-publication assignment. For determining the relevance of an identified gene-to-DN association, i.e. if a gene was reported significantly more frequent with DN as by chance, a statistics procedure was applied: Contingency tables were computed by splitting the full set of PubMed IDs first with respect to the occurrence in the context of DN and secondly with respect to the occurrence of each specific gene. 265 unique gene symbols identified as having enriched association with diabetic nephropathy could be retrieved. The same approach was applied for the headings “diabetic nephropathy” AND “biomarker”. In this restricted setting 108 unique gene symbols were identified.

2.3 Data set, clinical trials mining

Information on clinical trials focusing on DN was retrieved from http://www.clinicaltrials.gov maintained by the NIH. 260 studies could be identified by using the keyword “diabetic nephropathy” in the search field provided at the clinical trials web site. Among these studies 238 specifically focusing on type 2 diabetes were further considered. 111 interventional drugs were extracted from these 238 studies, and were subsequently queried in the DrugBank database [39]. 64 drugs could be identified in DrugBank holding 176 drug-gene relations, linking to 144 unique gene symbols.

2.4 Data set, patent mining

Patent Lens (http://www.patentlens.net) was used to identify molecular features provided in patents and patent applications associated with DN. Patent Lens offers the full text on granted patents from the US, Europe and Australia, as well as patent applications from the US, WIPO/PCT and Australia. A Patent Lens search was performed using the full-text keyword query “diabetic AND nephropathy”, providing 30,604 hits. Only patents linked to patent classes 424 (Drug, bio-affecting and body treating), 436 (Chemistry: Analytical and immunological testing), and 514 (extension to category 424) were retained. For retrieving only patents with evident association with DN a search on each patent was done for extracting patents where either “diabetic nephropathy” was present in the patent title or in the patent claims, or where the search string was found at least five times in the full-text, or the string was found in full text together with specific tags in the claims (“diabetic”, “diabetes” or “diab” AND “nephropathy”, “kidney” or “nephro”). This process resulted in a list of 1,561 patents with evident link to DN. From these the claims were kept for identification of genes and proteins associated with DN. The present list of valid NCBI symbols and full gene names was retrieved from the current version of GenBank, holding 19,066 genes, further expanded by associated 72,559 deprecated symbols and aliases. These were filtered by an English thesaurus (https://addons.mozilla.org/en-US/firefox/language-tools) containing 62,109 words, resulting in 18,905 valid symbols and full names, and 71,054 aliases. This query set was then applied for a case insensitive search among text tokens extracted from patent claims, resulting in 901 gene symbols.

2.5 Protein interaction network

All data sets generated on DN were mapped to gene symbols as common name space. These were subsequently mapped on a full human proteome interaction network (omicsNET, [40]). This network aims at a complete representation of protein coding genes, furthermore integrating the various types of protein interactions (being physical interactions, procedural interactions, or paralogs). Our network holds 26,419 nodes representing canonical proteins (as provided by SwissProt in combination with ENSEMBL). The starting point in the construction of omicsNET was the set of experimentally determined PPIs as obtained through the unification of available PPI sources (IntAct, OPHID, BioGrid, KEGG, PANTHER and Reactome). On top, data on tissue specific gene expression, sub-cellular localization (retrieved from

SwissProt and computed by WPSORT), ontology (GO molecular function and biological process) and pathway (KEGG, PANTHER) annotation and protein domains (PDB) were included. This procedure resulted in a maximum of seven data sources per node. On the basis of this parameterization we inferred the probability of a relation between node pairs by weighting the strength of the arguments in favor and against such a relation as encoded in the given parameters for a specific pair, technically represented by a metafunction. This procedure resulted in a matrix of pair relation weights being in the interval [-1,2], where negative values indicate lack of relation, and positive values indicate a relation for a given pair. The accuracy for estimating a relation between nodes certainly depends on the number of available parameters. E.g. tissue specific gene expression (represented in the metafunction as pair-wise correlation coefficient) is available only for 20,282 of the in total 26,419 nodes. The assumption therefore is that a higher number of valid parameters for a given pair (maximum seven parameters) increases the validity of a given edge weight, subsequently termed as evidence level of the relation. For 19,236 nodes at least four out of the seven parameters were found valid for computing an edge weight, and by selecting an edge weight cutoff of >= 0.58 a graph holding 18,948 nodes resulted, each specified by a unique molecular identifier. This node set and respective interaction network was then annotated by the diverse data sets, i.e. for each node information was added regarding identification in Omics studies, literature, clinical trials or patent text.

2.6 PPI network layout

For allowing an interpretation of the annotated omicsNET graph a functional layout in scope of KEGG pathways was performed utilizing the KEGG status as of October 2010. From the in total 214 pathways provided by KEGG all entries specifically focusing on disease phenotypes (as “pathways in cancer”) were removed, resulting in a set of 171 “generic” pathways. For each pathway, the KEGG identifiers of the assigned genes were retrieved and mapped to ENSEMBL protein identifiers via Entrez gene identifiers. The cross-referencing used in the process was obtained from the current version of the ENSEMBL database, resulting in 20,462 ENSEMBL protein identifiers. The set was then restricted to proteins present in our PPI network as described above, yielding 7,009 items. 2,924 of these were present in more than one KEGG pathway, and for assuring uniqueness of assignment each of these objects was assigned only to the KEGG pathway already holding the node which showed the strongest weight (as computed in omicsNET) to the non-uniquely assigned node. 2,644 nodes could be assigned this way, resulting in 151 populated pathways further considered. The same procedure was then applied for all nodes from the omicsNET node set not assigned at all to a KEGG pathway. This procedure allowed assigning the remaining nodes to a single pathway of the in total 151 pathways, where the allocation either rested on the given assignment by KEGG as such, or by the edge weights available in omicsNET. Of the 18,948 nodes provided in omicsNET in total 17,995 nodes could be assigned. This approach allowed the clustering of omicsNET in KEGG pathways, subsequently easing interpretation of the DN data sets. For analyzing this KEGG-based clustering of omicsNET a treemap [41] was computed. The inter-pathway similarity between two pathways was calculated as the average over all omicsNET edge weights between the two pathways. This similarity was then transformed into inter-pathway distance. On the basis of these distances a hierarchical clustering using Ward's linkage was computed and represented as a tree with merge steps as dummy nodes and pathways as leaves. From the resulting tree, a treemap was constructed using the Treemap program (http://www.cs.umd.edu/hcil/treemap-history) enabling an inspection of pathway distances. For evaluating if a pathway showed a significant enrichment of annotation with respect to a specific data category (Omics, literature, clinical trials, patents) a Fisher's Exact Test was applied using a p-value < 0.05 as significance level.

3 Results

3.1 Linking disease processes with biomarkers and targets

Due to the prevalence of DN a solid body of data is available, spanning various Omics screening levels and scientific literature, but also patent text and clinical trial descriptions. We extracted relevant data files and documents on a keyword search basis with the aim of identifying molecular features associated with the disease, and linking these into a canonical gene/protein interaction network (omicsNET)

All primary sources were processed for gene-disease associations, being directly provided from Omics screening results, via MeSH-disease-to-pubmed assignment coupled with gene-to-pubmed data for MEDLINE sources, via drug-gene assignment for clinical trials associated with DN, and via identification of gene symbols and names in DN-relevant patent text. For each extracted association the respective data structure of the molecular feature was expanded for mirroring the identified association, also providing the type of data source. In this approach no specific level of detail regarding the type of association was taken into consideration (e.g. direction and amplitude of differential regulation/abundance as derived in Omics profiling), but solely the fact of a qualified association between a molecular feature and the disease was retrieved. Next to this disease-specific annotation each molecular feature holds a second layer of annotation specifically used for computing a relation score between genes/proteins (molecular context). Applying this procedure to the full set of human protein coding genes provided a complete interaction network for the human proteome. Functional grouping of the molecular features in line with core KEGG pathways allows identification of pathways specifically enriched by one (or more) feature sets holding a disease annotation. Based on the sources used for feature retrieval qualitative layers can be defined holding i) a description of the pathophysiology of the disease (represented by features coming from Omics profiling and literature mining), and ii) holding features reported in the realm of pre-clinical or clinical application, i.e. identified explicitly in the context of biomarkers and therapy targets found in scientific publications, patent text and clinical trial descriptions (in the following denoted as “clinical context”). As a result of this integration strategy each affected pathway becomes amenable for analyzing its association to the pathophysiology as well as to biomarkers and therapy targets.

3.2 Direct feature overlap

First question to be addressed when analyzing heterogeneous feature lists is their direct overlap on the level of molecular identifiers. Table 2 provides an overview on the molecular feature sets derived from the various data and literature sources.

Insert Table 2 here

The source providing most hits is patent text with 901 gene symbols, followed by transcriptomics studies with 708 symbols. Mining of scientific literature provided 287 symbols linked with DN, and 108 symbols associated specifically with biomarkers in the context of DN. On the level of a direct NCBI gene symbol overlap 44 symbols are identified in both, literature and Omics profiles. The gene symbol overlap of combining Omics and literature (the pathophysiology set), and individually comparing this set with the other sources classified under clinical context is shown in Fig. 1A.

Insert Figure 1 here

Of the in total 1,094 unique symbols assigned to pathophysiology 102 are also found in the restricted literature mining focused on biomarkers, 144 are represented in patents, and 39 in clinical trial descriptions. Specific comparison of Omics tracks on a feature level is provided in Fig. 1B. From the in total 708 gene symbols reported from transcriptomics 5 are also reflected on the SNP, 5 on the proteome, and 6 on the metabolome level. Weak overlap on the level of individual features, however, has been reported in various meta-studies of Omics profiles [40, 42]. This alleged heterogeneity changes when going to the level of pathways populated by identified features instead of comparing individual features as such.

3.3 GSEA, KEGG pathway level

A gene set enrichment analysis (GSEA) [43] utilizing KEGG pathways provided a more coherent picture when comparing affected pathways grounded on data sets assigned to pathophysiology and to clinical context, as shown in Table 3.

Insert Table 3 here

Two pathways, namely the renin-angiotensin system (hsa04614) as well as the complement and coagulation cascades (hsa04610) appeared as significantly affected in both, the pathophysiology as well as the clinical context layer. Although the total number of features for both layers is in the very same range only five pathways were found to be significantly affected when mapping the pathophysiology feature set, but 28 pathways were found significantly affected when supplying the clinical context list. Among these are pathways frequently reported in the context of kidney disease, as MAPK signaling [44], VEGF signaling [45, 46], or TGF-beta signaling [47]. For completeness and as positive control we explicitly included the type II diabetes mellitus pathway also provided in KEGG, indeed showing enrichment for the clinical context data set.

3.4 GSEA, extended KEGG pathway level

For overcoming the limitation in coverage of genes in KEGG, but also for going beyond the procedural interactions provided in KEGG, we generated an extended KEGG pathway set holding in total 17,995 proteins. We mapped the five feature data sets (literature, combined Omics, literature with focus on biomarker, patents, and clinical trials) individually to the extended pathways and searched for pathways showing significant enrichment in affected features (i.e. being members of the DN data sets) utilizing a treemap representation for reflecting pathway neighborhood.

Treemaps align pathways with respect to their relatedness, but only a minor number of affected pathways were found in proximity for any of the five data sets. Furthermore, only three pathways were coherently found as enriched in all five maps, namely i) complement and coagulation cascade, ii) the renin-angiotensin system, and iii) the PPAR signaling pathway, where the first two were also identified on the KEGG level, and PPAR was in a first place only found on the level of pathophysiology. Only one additional pathway is jointly affected for Omics data and general literature data, namely “focal adhesion”. For the literature data set “TGF-beta signaling” is an affected and direct neighbor of the renin-angiotensin pathway, whereas for the Omics data set “vascular smooth muscle contraction” is significantly affected, apparently reflecting Ca^{2+} signaling. Two pathways were coherently identified on the clinical context level also being direct neighbors in the treemap representation, namely i) neuroactive receptor-ligand interactions and ii) cytokine-cytokine receptor interaction. The total number of pathways found as significantly enriched on the individual data set level, as well as the comparative display of affected pathways is provided in Table 4.

Insert Table 4 here

Among the 151 pathways represented in the treemaps at maximum 20 are affected by a single source (patent feature list), and only three to seven pathways are jointly found as enriched when comparing two data sets. In our pathway representation each pathway is populated by members assigned by KEGG, and members derived from the extended assignment of gene symbols based on omicsNET edge weights. On the basis of omicsNET all members of a specific pathway have relations assigned, enabling the extraction of pathway-specific subgraphs. The subgraph of the extended renin-angiotensin pathway is shown in Fig. 2.

Insert Figure 2 here

Of the 37 members (level of gene symbols) represented in the extended KEGG "renin-angiotensin pathway" (compared to 17 members of the original KEGG pathway hsa04614) 11 are found in at least one data set, with multiple data source annotation for the proteins ACE (angiotensin 1 converting enzyme 1), ACE2 (angiotensin 1 converting enzyme 2), AGT (angiotensin), CMA1 (chymase 1), ENPEP (glutamyl aminopeptidase), and REN (renin). Further features identified as relevant but not being members of the original KEGG pathway are CCK (Cholecystokinin), XPNPEP2 (X-prolyl aminopeptidase (aminopeptidase P) 2, DPEP (a renal dipeptidase), and KLK7 (kallikrein 7).

Three pathways were found individually enriched by all five extracted data sets. Biomarkers (derived from the focused literature mining) and therapy targets (as derived from clinical trials text) assigned to these pathways are listed in Table 5.

Insert Table 5 here

Biomarkers and targets associated with hsa04614 (renin-angiotensin system) obviously hold ACE and AGT on both, biomarker and target level. A further consensus finding were the pathways has04080 (neuroactive receptor-ligand interactions) and hsa04060 (cytokine-cytokine receptor interaction), being enriched in the clinical context setting resting on the focused biomarker search in the scientific literature, patent information, as well as clinical trial text. In total 16 specific drugs are linked to hsa04080 (mainly sartans acting as angiotensin II receptor antagonists), and two to hsa04060 (statins for lowering cholesterol levels via inhibiting HMG-CoA reductase) according to the gene-drug assignments from DrugBank. Their main associated disease indications as assigned in MeSH are given in Table 6, clearly reflecting cardiovascular medication for drugs linked to hsa04080 and cholesterol reduction, and lipid balance and diabetes for hsa04060.

Insert Table 6 here

MeSH terms associated with PubMed hits found by individually searching these 16 drugs together with a provided MeSH modifier for obtaining systematic reviews of value to clinicians ("systematic[sb]") allowed us to count and rank literature occurrence of diseases associated with these drugs, as listed in Table 7.

Insert Table 7 here

4 Discussion

Omics technologies have significantly broadened our experimental capabilities for describing the molecular status of a given biological matrix (tissue, blood or urine level). Improvements in experimental workup of samples, SOPs for Omics screening, and standardization in reporting including both, experimental description as well as execution have provided the ground for utilizing Omics also in the clinical context. Here individual Omics tracks, as proteomics in the field of diabetic nephropathy, have already entered validation in disease diagnosis and prognosis. Workup of microdissected kidney tissue forwarded to transcriptomics has resulted in the discovery of major pathways in the tubular compartments seen with chronic kidney disease [46]. However, most of these procedures have been implemented “within the Omics domain”, still centrally driven by statistics procedures aimed at identifying differentially regulated transcripts or differentially abundant proteins. This domain separation also becomes evident by the highly valuable Omics profile consolidation efforts e.g. seen for transcriptomics at ArrayExpress [48], or for proteomics with PRIDE [49], or more general gene-centric data consolidation as with GeneCards [50]. Disease-specific Omics repositories are unfortunately still sparse and valuable examples as Oncomine (Oncomine™ - Compendia Bioscience, Ann Arbor, MI; focus on neoplasm) or Nephromine (www.nephromine.org, focus on renal expression profiles) are transcriptomics centered. Another initiative, SysKid (www.syskid.eu), takes a different route, namely limiting the clinical phenotype under analysis, but in turn broadening the Omics disease annotation beyond transcriptomics by also including GWAS, proteomics and metabolomics. The present paper follows this concept for characterizing diabetic nephropathy, furthermore including text mining results from scientific literature, patent text as well as clinical trial descriptions yielding 1,094 features characterizing the pathophysiology, and 1,059 features associated with the clinical context of the disease.

A first finding in cross-Omics data comparison is on the level of direct feature comparison, where meta-analysis within an Omics domain [40] as well as cross Omics domains [27] shows sparse overlap. Utilizing this procedure for the given Omics data sets on DN provides the same conclusion, and also when broadening the features included beyond Omics does not substantially change this picture. In terms of biological causality with respect to abundance levels from Omics (the central result from a statistics-driven profile analysis), however, this finding is not surprising: SNP for instance may affect efficacy of a protein’s function but this fact is eventually not mirrored on the transcript level with respect to different concentration as determined in a microarray experiment. Furthermore, different Omics profiling done in different sample types (e.g. tissue, plasma, urine) characterize joint, but also up- and downstream processes. Based on this background cross-Omics analysis on the level of networks and pathways became the method of choice with the underlying assumption that the heterogeneity of identified individual features consolidates on the level of functional units (pathways), as these provide a causal link between concerted molecular processes being responsible for a given disease phenotype. For pathways also interdependencies can be described (e.g. encoded as pathway distance in treemaps), which might support elucidation of sequences of processes (as inherently given when e.g. integrating tissue transcriptomics and urinary proteomics profiles). Data on functional units, represented as interaction networks (graphs), is available for a number of generic cellular processes with KEGG as the most prominent example. Mapping the data sets on DN to KEGG and performing a GSEA identified five pathways as affected when using combined Omics and literature feature sets. 80 out of the in total 1,094 features are assigned to these five pathways, all other features are assigned to other pathways where the statistical analysis of the number of total features given in a pathway and the number of respective features also found in either the Omics or the literature data set did not show significant enrichment. In any case, identification of affected pathways is significantly improved when combining the different Omics domains. The number of enriched pathways is substantially expanded when using the clinical context lists, resulting in 28 pathways resting on in total 406 features from the 1,059 features assigned to clinical context. Two pathways are found as affected for both, pathophysiology as well as clinical context, namely the renin-angiotensin system as well as the complement and coagulation cascade. The renin-angiotensin system depicts an important mechanism in the regulation of blood pressure with an increased production of renin in the kidney leading to a constriction of blood vessels and an increased blood pressure. Next to the direct inhibition of renin in order to lower blood pressure in patients with diabetic nephropathy, therapeutic options targeting the angiotensin-converting enzyme or the angiotensin II receptor are in clinical use. Next to the blood pressure lowering capabilities of angiotensin-converting enzyme inhibitors (ACEIs) and angiotensin II type 1 receptor blockers (ARBs), their independent anti-proteinuric effect made them a major factor in treatment of chronic kidney disease. The complement cascade is an essential part of the innate immune system, and the complement components C5b-9 have also been detected in urine of diabetic nephropathy patients [51]. The link between complement and the coagulation cascade and the contribution to inflammation and other diseases has been outlined by Markiewski and colleagues [52].

However, available pathway data as encoded in KEGG are far from being complete with respect to the number of protein coding genes, and fall short in integrating the various types of interactions. For expanding both, comprehensiveness as well as coverage of interaction types we utilized omicsNET aimed at a complete representation of protein coding genes in a network of relations encapsulating various types of interactions. However, omicsNET does not provide functional units, and recalling the above said such units are deemed necessary for consolidating and interpreting the diverse Omics data. Therefore nodes not represented in KEGG were assigned to KEGG pathways using the omicsNET relations weight as assignment criterion: Each node in the first place not represented in KEGG was assigned to a given KEGG node showing strongest relation with the non-assigned node. Result of this procedure is a clustering of omicsNET in KEGG pathway categories.

The diverse data sets retrieved on DN were mapped on this extended KEGG category set, and GSEA was again performed resulting in 10 to 20 enriched pathways for each data source. The two pathways already seen in KEGG, namely the renin-angiotensin as well as the complement and coagulation cascade, were again identified, but in this setting significant enrichment was given individually for all five data sets. Due to the fact that omicsNET is underlying the pathways, extraction of the pathway specific relations network became possible. The functional unit presents as a single connected component, and short paths are seen for major players of this pathway including ACE, REN and AGT, as well as of members which were assigned to this pathway as KLK7, linking the kallikrein-kinin system. Due to mapping all data sources on a common name space the assignment of features to biomarkers and therapy targets is straight forward. Cholesterol lowering drugs, insulin sensitizers, and ACE inhibitors are prominently linked in this category together with sulodexide, a glycosaminoglycan mixture actively tested in DN patients with urinary albumin excretion [53]. Additionally the PPAR signaling pathway, in the first setup only seen for the combined Omics and literature data set, became significant for all five data sets. PPARs are involved in modulating insulin resistance, hypertension, dyslipidemia, obesity, and inflammation. PPARs depict promising alternative therapeutic targets next to the above mentioned molecules of the renin-angiotensin pathway for diseases like type 2 diabetes, obesity, hypertension, hyperlipidemia, or atherosclerosis. A number of clinical trials suggest the renoprotective effects of PPAR agonists [54].

From this unbiased data selection and analysis three pathways became evident allowing a conclusive link of biomarkers and therapy targets on the basis of a molecular description of the pathophysiology of the disease. However, two additional pathways, functionally being in close proximity, were identified congruently for the clinical context data sets, but not being significantly affected on the pathophysiology level, namely “neuroactive receptor-ligand interactions” and “cytokine-cytokine receptor interaction”. Next to statins found for the cytokine-cytokine receptor interaction pathway the class of sartans (angiotensin II receptor antagonists) is prominently linked to the neuroactive receptor-ligand interactions. Cross-evaluating the drugs found in these two pathways with assigned disease names according to NCBI MeSH clearly reflects hypertension and diseases afflicted with lipid metabolism.

The molecular pathways retrieved by the multi-source consolidation perfectly match the clinical view regarding risk factors for developing DN, involving among others hyperglycemia and cytokine activation, and associated therapy regimes. Data integration also demonstrates the intricate entanglement of decreased kidney function and hypertension in the realm of diabetes mellitus. Certainly the various data sets utilized in this work carry different levels of evidence and specific biases. For explorative (bias-free) Omics studies the study design, and here specifically proper case and control definition as well as appropriate statistical power, define evidence of features termed as relevant in the context of DN. Automated data retrieval from literature and patents on the other hand results in less evident and also biased feature extraction. Integration of the clinical relevance data space with Omics profiles nevertheless is supportive for interpretation of explorative data in the realm of known associations next to identification of novel features and respective pathways. Consolidation of molecular features on interaction networks provides furthermore the basis for linking molecular profiles with associated biomarkers and therapy targets. Expanding the concept presented here towards also linking detailed clinical data for each individual Omics profile as yet another layer would generate patient (cohort) specific molecular pathology landscapes, which in a next step could be linked with cohort specific diagnostics and therapy regimes. An apparent shortcoming of the approach presented here is the on a clinical level still broad spectrum of DN (level of albuminuria, stage of the disease), which was not specifically addressed in the course of data set retrieval. For implementing a true Systems Biology approach, however, specific clinical data specifying the context of Omics profiles, but also specifying the context of individual molecular features are necessary. Feeding a multilayered reference network as delineated in this work with patient specific Omics profiles/marker profiles, and analyzing affected pathways in the realm of specific clinical data might then offer a route towards individualized therapy strategies.

Acknowledgements:

The research leading to these results has received funding from the European Community's Seventh Framework Programme under grant agreement n° HEALTH –F2–2009–241544.

Conflict of interest:

The authors declare no conflict of interest.

5 References

[1] Newman, D. J., Mattock, M. B., Dawnay, A. B., Kerry, S.*, et al.*, Systematic review on urine albumin testing for early detection of diabetic complications. *Health Technol Assess* 2005, *9*, iii-vi, xiii-163.
[2] Bojestig, M., Arnqvist, H. J., Hermansson, G., Karlberg, B. E., Ludvigsson, J., Declining incidence of nephropathy in insulin-dependent diabetes mellitus. *N Engl J Med* 1994, *330*, 15-18.
[3] Finne, P., Reunanen, A., Stenman, S., Groop, P. H., Gronhagen-Riska, C., Incidence of end-stage renal disease in patients with type 1 diabetes. *Jama* 2005, *294*, 1782-1787.
[4] Adler, A. I., Stevens, R. J., Manley, S. E., Bilous, R. W.*, et al.*, Development and progression of nephropathy in type 2 diabetes: the United Kingdom Prospective Diabetes Study (UKPDS 64). *Kidney Int* 2003, *63*, 225-232.
[5] Wild, S., Roglic, G., Green, A., Sicree, R., King, H., Global prevalence of diabetes: estimates for the year 2000 and projections for 2030. *Diabetes Care* 2004, *27*, 1047-1053.
[6] Ozyilmaz, A., Bakker, S. J., de Zeeuw, D., de Jong, P. E., Gansevoort, R. T., Selection on albuminuria enhances the efficacy of screening for cardiovascular risk factors. *Nephrol Dial Transplant* 2010, *25*, 3560-3568.
[7] Mann, J. F., Schmieder, R. E., McQueen, M., Dyal, L.*, et al.*, Renal outcomes with telmisartan, ramipril, or both, in people at high vascular risk (the ONTARGET study): a multicentre, randomised, double-blind, controlled trial. *Lancet* 2008, *372*, 547-553.
[8] Bakris, G. L., Sarafidis, P. A., Weir, M. R., Dahlof, B.*, et al.*, Renal outcomes with different fixed-dose combination therapies in patients with hypertension at high risk for cardiovascular events

(ACCOMPLISH): a prespecified secondary analysis of a randomised controlled trial. *Lancet* 2010, *375*, 1173-1181.

[9] Levey, A. S., Cattran, D., Friedman, A., Miller, W. G.*, et al.*, Proteinuria as a surrogate outcome in CKD: report of a scientific workshop sponsored by the National Kidney Foundation and the US Food and Drug Administration. *Am J Kidney Dis* 2009, *54*, 205-226.

[10] Blezquez-Medela, A. M., Lopez-Novoa, J. M., Martinez-Salgado, C., Mechanisms involved in the genesis of diabetic nephropathy. *Curr Diabetes Rev* 2010, *6*, 68-87.

[11] Cushman, W. C., Evans, G. W., Byington, R. P., Goff, D. C., Jr.*, et al.*, Effects of intensive blood-pressure control in type 2 diabetes mellitus. *N Engl J Med* 2010, *362*, 1575-1585.

[12] Fioretto, P., Steffes, M. W., Sutherland, D. E., Goetz, F. C., Mauer, M., Reversal of lesions of diabetic nephropathy after pancreas transplantation. *N Engl J Med* 1998, *339*, 69-75.

[13] Holman, R. R., Paul, S. K., Bethel, M. A., Matthews, D. R., Neil, H. A., 10-year follow-up of intensive glucose control in type 2 diabetes. *N Engl J Med* 2008, *359*, 1577-1589.

[14] Holman, R. R., Paul, S. K., Bethel, M. A., Neil, H. A., Matthews, D. R., Long-term follow-up after tight control of blood pressure in type 2 diabetes. *N Engl J Med* 2008, *359*, 1565-1576.

[15] KDOQI, Clinical Practice Guidelines and Clinical Practice Recommendations for Diabetes and Chronic Kidney Disease. www.kdoqi.org.

[16] Pan, Y., Guo, L. L., Jin, H. M., Low-protein diet for diabetic nephropathy: a meta-analysis of randomized controlled trials. *Am J Clin Nutr* 2008, *88*, 660-666.

[17] Gaede, P., Vedel, P., Larsen, N., Jensen, G. V.*, et al.*, Multifactorial intervention and cardiovascular disease in patients with type 2 diabetes. *N Engl J Med* 2003, *348*, 383-393.

[18] Howard, B. V., Roman, M. J., Devereux, R. B., Fleg, J. L.*, et al.*, Effect of lower targets for blood pressure and LDL cholesterol on atherosclerosis in diabetes: the SANDS randomized trial. *Jama* 2008, *299*, 1678-1689.

[19] Kottgen, A., Pattaro, C., Boger, C. A., Fuchsberger, C.*, et al.*, New loci associated with kidney function and chronic kidney disease. *Nat Genet* 2010, *42*, 376-384.

[20] Cohen, C. D., Lindenmeyer, M. T., Eichinger, F., Hahn, A.*, et al.*, Improved elucidation of biological processes linked to diabetic nephropathy by single probe-based microarray data analysis. *PLoS One* 2008, *3*, e2937.

[21] Merchant, M. L., Klein, J. B., Proteomics and diabetic nephropathy. *Curr Diab Rep* 2005, *5*, 464-469.

[22] Sebedio, J. L., Pujos-Guillot, E., Ferrara, M., Metabolomics in evaluation of glucose disorders. *Curr Opin Clin Nutr Metab Care* 2009, *12*, 412-418.

[23] Rossing, K., Mischak, H., Dakna, M., Zurbig, P.*, et al.*, Urinary proteomics in diabetes and CKD. *J Am Soc Nephrol* 2008, *19*, 1283-1290.

[24] Alkhalaf, A., Zurbig, P., Bakker, S. J., Bilo, H. J.*, et al.*, Multicentric validation of proteomic biomarkers in urine specific for diabetic nephropathy. *PLoS One* 2010, *5*, e13421.

[25] Fox, C. S., Gona, P., Larson, M. G., Selhub, J.*, et al.*, A Multi-Marker Approach to Predict Incident CKD and Microalbuminuria. *J Am Soc Nephrol* 2010, *21*, 2143-2149.

[26] Molina, F., Dehmer, M., Perco, P., Graber, A.*, et al.*, Systems biology: opening new avenues in clinical research. *Nephrol Dial Transplant* 2010, *25*, 1015-1018.

[27] Mühlberger, I., Mönks, K., Bernthaler, A., Jandrasits, C.*, et al.*, Integrative bioinformatics analysis of proteins associated with the cardiorenal syndrome. *Int J Nephrol* 2011, *2011*, 10.

[28] Perco, P., Wilflingseder, J., Bernthaler, A., Wiesinger, M.*, et al.*, Biomarker candidates for cardiovascular disease and bone metabolism disorders in chronic kidney disease: a systems biology perspective. *J Cell Mol Med* 2008, *12*, 1177-1187.

[29] Lusis, A. J., Weiss, J. N., Cardiovascular networks: systems-based approaches to cardiovascular disease. *Circulation* 2010, *121*, 157-170.

[30] Hidalgo, C. A., Blumm, N., Barabasi, A. L., Christakis, N. A., A dynamic network approach for the study of human phenotypes. *PLoS Comput Biol* 2009, *5*, e1000353.
[31] Barrenas, F., Chavali, S., Holme, P., Mobini, R., Benson, M., Network properties of complex human disease genes identified through genome-wide association studies. *PLoS One* 2009, *4*, e8090.
[32] Dezso, Z., Nikolsky, Y., Nikolskaya, T., Miller, J.*, et al.*, Identifying disease-specific genes based on their topological significance in protein networks. *BMC Syst Biol* 2009, *3*, 36.
[33] Dudley, J. T., Butte, A. J., Identification of discriminating biomarkers for human disease using integrative network biology. *Pac Symp Biocomput* 2009, 27-38.
[34] Goh, K. I., Cusick, M. E., Valle, D., Childs, B.*, et al.*, The human disease network. *Proc Natl Acad Sci U S A* 2007, *104*, 8685-8690.
[35] Ozgur, A., Vu, T., Erkan, G., Radev, D. R., Identifying gene-disease associations using centrality on a literature mined gene-interaction network. *Bioinformatics* 2008, *24*, i277-285.
[36] Xu, M., Kao, M. C., Nunez-Iglesias, J., Nevins, J. R.*, et al.*, An integrative approach to characterize disease-specific pathways and their coordination: a case study in cancer. *BMC Genomics* 2008, *9 Suppl 1*, S12.
[37] Li, J., Zhu, X., Chen, J. Y., Building disease-specific drug-protein connectivity maps from molecular interaction networks and PubMed abstracts. *PLoS Comput Biol* 2009, *5*, e1000450.
[38] Wishart, D. S., Knox, C., Guo, A. C., Eisner, R.*, et al.*, HMDB: a knowledgebase for the human metabolome. *Nucleic Acids Res* 2009, *37*, D603-610.
[39] Wishart, D. S., Knox, C., Guo, A. C., Cheng, D.*, et al.*, DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res* 2008, *36*, D901-906.
[40] Bernthaler, A., Mühlberger, I., Fechete, R., Perco, P.*, et al.*, A dependency graph approach for the analysis of differential gene expression profiles. *Mol Biosyst* 2009, *5*, 1720-1731.
[41] Johnson, B., Shneiderman, B., *Proceedings of the 2nd conference on Visualization '91*, IEEE Computer Society Press, San Diego, California, 1991.
[42] Rapberger, R., Perco, P., Sax, C., Pangerl, T.*, et al.*, Linking the ovarian cancer transcriptome and immunome. *BMC Syst Biol* 2008, *2*, 2.
[43] Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S.*, et al.*, Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 2005, *102*, 15545-15550.
[44] Sengupta, U., Ukil, S., Dimitrova, N., Agrawal, S., Expression-based network biology identifies alteration in key regulatory pathways of type 2 diabetes and associated risk/complications. *PLoS One* 2009, *4*, e8100.
[45] Mironidou-Tzouveleki, M., Tsartsalis, S., Tomos, C., Vascular Endothelial Growth Factor (VEGF) in the Pathogenesis of Diabetic Nephropathy of Type 1 Diabetes Mellitus. *Curr Drug Targets* 2011, *12*, 107-114.
[46] Rudnicki, M., Perco, P., Enrich, J., Eder, S.*, et al.*, Hypoxia response and VEGF-A expression in human proximal tubular epithelial cells in stable and progressive renal disease. *Lab Invest* 2009, *89*, 337-346.
[47] Jiang, W., Zhang, Y., Wu, H., Zhang, X.*, et al.*, Role of cross-talk between the Smad2 and MAPK pathways in TGF-beta1-induced collagen IV expression in mesangial cells. *Int J Mol Med* 2010, *26*, 571-576.
[48] Parkinson, H., Kapushesky, M., Kolesnikov, N., Rustici, G.*, et al.*, ArrayExpress update--from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Res* 2009, *37*, D868-872.
[49] Vizcaino, J. A., Cote, R., Reisinger, F., Foster, J. M.*, et al.*, A guide to the Proteomics Identifications Database proteomics data repository. *Proteomics* 2009, *9*, 4276-4283.
[50] Safran, M., Dalah, I., Alexander, J., Rosen, N.*, et al.*, GeneCards Version 3: the human gene integrator. *Database (Oxford)*, *2010*, baq020.

[51] Morita, Y., Ikeguchi, H., Nakamura, J., Hotta, N.*, et al.*, Complement activation products in the urine from proteinuric patients. *J Am Soc Nephrol* 2000, *11*, 700-707.
[52] Markiewski, M. M., Nilsson, B., Ekdahl, K. N., Mollnes, T. E., Lambris, J. D., Complement and coagulation: strangers or partners in crime? *Trends Immunol* 2007, *28*, 184-192.
[53] Lambers Heerspink, H. J., Fowler, M. J., Volgi, J., Reutens, A. T.*, et al.*, Rationale for and study design of the sulodexide trials in Type 2 diabetic, hypertensive patients with microalbuminuria or overt nephropathy. *Diabet Med* 2007, *24*, 1290-1295.
[54] Kume, S., Uzu, T., Isshiki, K., Koya, D., Peroxisome proliferator-activated receptors in diabetic nephropathy. *PPAR Res* 2008, *2008*, 879523.

Tables

Table 1: Omics studies extracted from the literature.

study no.	Omics type	reference
1	transcriptomics	Baelde, H.J., Eikmans, M., Doran, P.P., Lappin, D.W., et al., Gene expression profiling in glomeruli from human kidneys with diabetic nephropathy. *Am J Kidney Dis* 2004, *43*, 636-650.
2	transcriptomics	Berthier, C.C., Zhang, H., Schin, M., Henger, A., et al., Enhanced expression of Janus kinase-signal transducer and activator of transcription pathway members in human diabetic nephropathy. *Diabetes* 2009, *58*, 469-477.
3	transcriptomics	Cohen, C.D., Lindenmeyer, M.T., Eichinger, F., Hahn, A., et al. Improved elucidation of biological processes linked to diabetic nephropathy by single probe-based microarray data analysis. *PLoS One* 2008, *3*, e2937.
4	transcriptomics	Rudnicki, M., Eder, S., Perco, P., Enrich, J., et al., Gene expression profiles of human proximal tubular epithelial cells in proteinuric nephropathies. *Kidney Int* 2007, *71*, 325-335.
5	proteomics	Dihazi, H., Müller, G.A., Lindner, S., Meyer, M., et al., Characterization of diabetic nephropathy by urinary proteomic analysis: identification of a processed ubiquitin form as a differentially excreted protein in diabetic nephropathy patients. *Clin Chem* 2007, *53*, 1636-1645.
6	proteomics	Jain, S., Rajput, A., Kumar, Y., Uppuluri, N., et al., Proteomic analysis of urinary protein markers for accurate prediction of diabetic kidney disorder. *J Assoc Physicians India* 2005, *53*, 513-520.
7	proteomics	Kim, H.-J., Cho, E.-H., Yoo, J.-H., Kim, P.-K., Shin, J.-S., Kim, M.-R., et al. (2007). Proteome analysis of serum from type 2 diabetics with nephropathy. Journal of proteome research, 6(2), 735-43.
8	proteomics	Mischak, H., Kaiser, T., Walden, M., Hillmann, M., et al., Proteomic analysis for the assessment of diabetic renal damage in humans. *Clin Sci (London)* 2004, *107*, 485-495.
9	proteomics	Otu, H.H., Can, H., Spentzos, D., Nelson, R.G., et al., Prediction of diabetic nephropathy using urine proteomic profiling 10 years prior to development of nephropathy. *Diabetes Care* 2007, 30, 638-643.
10	proteomics	Rossing, K., Mischak, H., Dakna, M., Zürbig, P., et al., Urinary proteomics in diabetes and CKD. *J Am Soc Nephrol* 2008, *19*, 1283-1290.
11	proteomics	Sharma, K., Lee, S., Han, S., Lee, S., et al., Two-dimensional fluorescence difference gel electrophoresis analysis of the urine proteome in human diabetic nephropathy. *Proteomics* 2005, *5*, 2648-2655.
12	metabolomics	Xia, J.F., Liang, Q.L., Liang, X.P., Wang, Y.M., et al., Ultraviolet and tandem mass spectrometry for simultaneous quantification of 21 pivotal metabolites in plasma from patients with diabetic nephropathy. *J Chromatogr B Analyt Technol Biomed*

		Life Sci 2009, *877*, 1930-1936.
13	metabolomics	Zhang, J., Yan, L., Chen, W., Lin, L., et al., Metabonomics research of diabetic nephropathy and type 2 diabetes mellitus based on UPLC-oaTOF-MS system. *Anal Chim Acta* 2009, *650*, 16-22.
14	SNP	Chambers, J.C., Zhang, W., Lord, G.M., van der Harst, P., et al., Genetic loci influencing kidney function and chronic kidney disease. *Nat Genet* 2010, *42*, 373-375.
15	SNP	Köttgen, A., Glazer, N.L., Dehghan, A., Hwang, S.J., et al., Multiple loci associated with indices of renal function and chronic kidney disease. *Nat Genet* 2009, *41*, 712-717.
16	SNP	Köttgen, A., Pattaro, C., Böger, C.A., Fuchsberger, C., et al., New loci associated with kidney function and chronic kidney disease. *Nat Genet* 2010, *42*, 376-384.
17	SNP	Lim, S.C., Liu, J.J., Low, H.Q., Morgenthaler, N.G., et al., Microarray analysis of multiple candidate genes and associated plasma proteins for nephropathy secondary to type 2 diabetes among Chinese individuals. Diabetologia. 2009, *52*, 1343-1351.

Table 1 lists the Omics type and scientific reference of public domain Omics sources used in this work.

Table 2: Feature extraction for DN.

level	source	classification	# features
pathophysiology	literature	“Diabetic Nephropathy”	287
	Omics	Transcriptomics	708
		Proteomics	30
		SNP	36
		Metabolomics	94
	total unique features		**1094**
clinical context	literature	“Diabetic Nephropathy” AND “Biomarker”	108
	trials	“Diabetic Nephropathy”	144
	patents	“Diabetic Nephropathy”	901
	total unique features		**1059**

Table 2 provides the mined sources assigned to either describing the pathophysiology of DN or to the clinical context of DN, source classification and keywords, and number of features (NCBI gene symbols) extracted.

Table 3: KEGG-based GSEA for features assigned to “pathophysiology” and “clinical context”.

KEGG term	# features	CC, hits	CC, p-value	PP, hits	PP, p-value
hsa04614:Renin-angiotensin system	17	10	0.01	10	0.00
hsa04610:Complement and coagulation cascades	69	31	0.00	19	0.01
hsa00760:Nicotinate and nicotinamide metabolism	24	-	-	10	0.04
hsa03320:PPAR signaling pathway	69	-	-	19	0.01
hsa00563:Glycosylphosphatidylinositol(GPI)-anchor biosynthesis	25	-	-	22	0.00
hsa04060:Cytokine-cytokine receptor interaction	262	93	0.00	-	-
hsa04010:MAPK signaling pathway	267	77	0.00	-	-
hsa04660:T cell receptor signaling pathway	108	43	0.00	-	-
hsa04620:Toll-like receptor signaling pathway	101	41	0.00	-	-
hsa04340:Hedgehog signaling pathway	56	29	0.00	-	-
hsa04930:Type II diabetes mellitus	47	25	0.00	-	-
hsa04621:NOD-like receptor signaling pathway	62	29	0.00	-	-
hsa04350:TGF-beta signaling pathway	87	35	0.00	-	-
hsa04722:Neurotrophin signaling pathway	124	42	0.00	-	-
hsa04062:Chemokine signaling pathway	187	54	0.00	-	-
hsa04672:Intestinal immune network for IgA production	49	24	0.00	-	-
hsa04916:Melanogenesis	99	36	0.00	-	-
hsa04012:ErbB signaling pathway	87	32	0.00	-	-
hsa04920:Adipocytokine signaling pathway	67	26	0.00	-	-
hsa04630:Jak-STAT signaling pathway	155	44	0.00	-	-
hsa04664:Fc epsilon RI signaling pathway	78	28	0.00	-	-
hsa04662:B cell receptor signaling pathway	75	27	0.00	-	-
hsa04910:Insulin signaling pathway	135	39	0.00	-	-
hsa04912:GnRH signaling pathway	98	30	0.00	-	-
hsa04020:Calcium signaling pathway	176	43	0.00	-	-
hsa04914:Progesterone-mediated oocyte maturation	86	26	0.01	-	-
hsa04520:Adherens junction	77	24	0.01	-	-
hsa04370:VEGF signaling pathway	75	23	0.01	-	-
hsa04080: Neuroactive ligand-receptor interaction	256	54	0.02	-	-
hsa04110: Cell cycle	125	32	0.02	-	-
hsa04150: mTor signaling pathway	52	18	0.02	-	-

Table 3 provides the KEGG identifier and term name, the total number of features assigned to this pathway, the number of features found to be affected utilizing ‘pathophysiology’ (PP) and ‘clinical context’ (CC) data sets, as well as associated p-values indicating the significance of association.

Table 4: Number of enriched pathways in extended KEGG.

source	Omics	literature	literature, biomarker	patents	clinical trials
Omics	**19**	6	3	5	6
literature		**17**	7	7	4
literature, biomarker			**10**	7	5
patents				**20**	7
clinical trials					**14**

Table 4 provides the feature source, the number of pathways enriched on the level of an individual source, and the number of pathways jointly affected in a pair-wise comparison.

Table 5: Biomarkers and therapy targets, consensus pathways.

extended KEGG pathway code	biomarker gene symbol	therapy target gene symbol	drug name
hsa04610	CD59	SERPINE1	Atorvastatin
	F3	F10, F2, SERPINC1	Enoxaparin
	FGB	C1QA, C1QB, C1QC, C1R, C1S	Rituximab
	HP	F2, MMP3, SERPINE1	Simvastatin
	KLKB1	SERPINE1, SERPINC1, SERPIND1	Sulodexide
	LPA		
	PTX3		
hsa03320	ALB	APOB, PON1, PPARA, PPARG	Atorvastatin
	APOB	ALB	Captopril, Insulin-Glargine
	FABP1	VDR	Paricalcitol
	HPR	PPARG	Pioglitazone, Rosiglitazone
	PON1	PPARA	Simvastatin
hsa04614	ACE	AGT, REN	Aliskiren
	ACE2	AGT	Atorvastatin, Simvastatin, Lisinopril, Irbesartan
	AGT	ACE	Benazepril, Captopril, Enalapril, Fosinopril, Lisinopril, Ramipril, Trandolapril

Biomarkers and targets including assigned drugs as represented in the extended KEGG pathways hsa04610 (complement and coagulation cascade), hsa03320 (PPAR signaling pathway) and hsa04614 (renin-angiotensin system).

Table 6: Drugs assigned to pathways consensually affected in the clinical context.

pathway	drug name	pharmacological action
hsa04080	ATENOLOL	Adrenergic beta-Antagonists, Anti-Arrhythmia Agents, Antihypertensive Agents, Sympatholytics
	BOSENTAN	Antihypertensive Agents
	CANDESARTAN	Angiotensin II Type 1 Receptor Blockers, Antihypertensive Agents
	CLONIDINE	Adrenergic alpha-Agonists, Analgesics, Antihypertensive Agents, Sympatholytics
	CORTICOTROPIN	Hormones
	DOXAZOSIN	Adrenergic alpha-Antagonists, Antihypertensive Agents
	EXENATIDE	Hypoglycemic Agents
	FOLIC ACID	Hematinics, Vitamin B Complex
	IRBESARTAN	Angiotensin II Type 1 Receptor Blockers, Antihypertensive Agents
	LOSARTAN	Angiotensin II Type 1 Receptor Blockers, Anti-Arrhythmia Agents, Antihypertensive Agents
	OLMESARTAN	Angiotensin II Type 1 Receptor Blockers
	PERINDOPRIL	Angiotensin-Converting Enzyme Inhibitors, Antihypertensive Agents
	SERTRALINE	Antidepressive Agents, Serotonin Uptake Inhibitors
	SPIRONOLACTONE	Aldosterone Antagonists, Diuretics
	TELMISARTAN	Angiotensin II Type 1 Receptor Blockers, Angiotensin-Converting Enzyme Inhibitors
	VALSARTAN	Angiotensin II Type 1 Receptor Blockers, Antihypertensive Agents
hsa04060	ATORVASTATIN	Anticholesteremic Agents, Hydroxymethylglutaryl-CoA Reductase Inhibitors
	SIMVASTATIN	Hydroxymethylglutaryl-CoA Reductase Inhibitors, Hypolipidemic Agents

Drugs assigned to the pathway hsa04080 (neuroactive receptor-ligand interactions) and hsa04060 (cytokine-cytokine receptor interaction), and pharmacological action.

Table 7: Indications assigned to drugs associated with consensual pathways given in the clinical context.

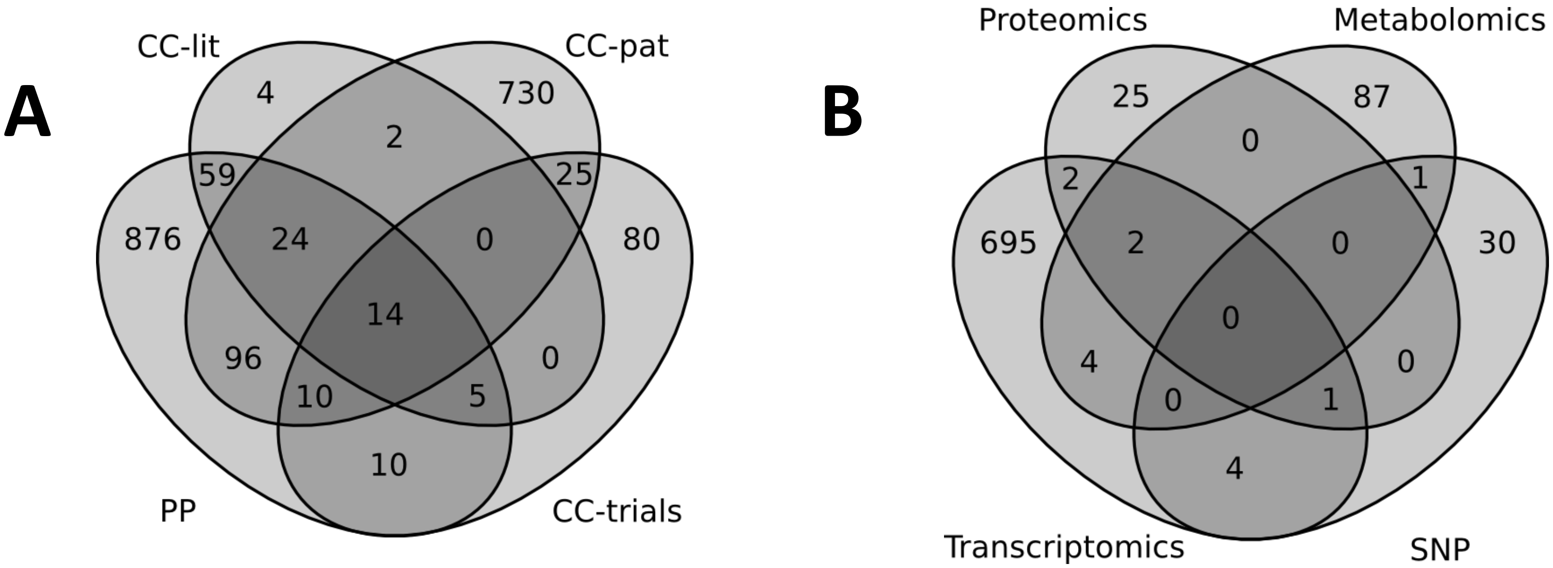
hsa04080	count	hsa0460	count
HYPERTENSION	68	HYPERCHOLESTEROLEMIA	38
COLORECTAL NEOPLASMS	56	CARDIOVASCULAR DISEASES	39
NEURAL TUBE DEFECTS	41	HYPERTENSION	5
CARDIOVASCULAR DISEASES	64	DIABETES MELLITUS	11
HEART FAILURE	24	RHABDOMYOLYSIS	4
FOLIC ACID DEFICIENCY/ HYPERHOMOCYSTEINEMIA	26	DEMENTIA	2
PREGNANCY COMPLICATIONS	13	STROKE	2
STROKE	13		
OPIOID-RELATED DISORDERS	13		
DIABETES MELLITUS, TYPE 2	12		

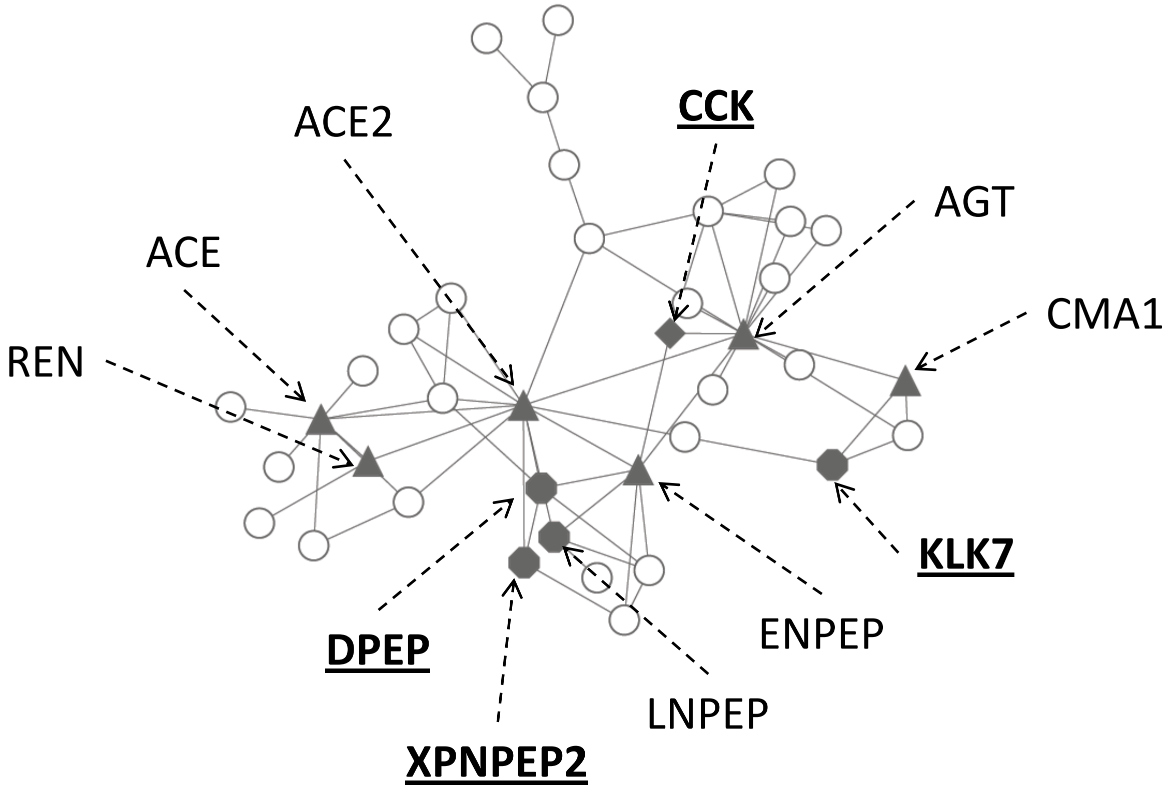
Name and frequency (count) of indications assigned to drugs found as associated with the pathway hsa04080 (neuroactive receptor-ligand interactions) and hsa04060 (cytokine-cytokine receptor interaction).

Figure captions

Figure 1: Venn diagrams comparing **A:** features assigned to pathophysiology (PP, Omics, literature) and their overlap with features extracted from a biomarker-biased literature search (CC-lit), patent text (CC-pat), and clinical trials (CC-trials) text. **B:** features assigned to the individual Omics tracks (SNP, transcriptomics, proteomics, metabolomics).

Figure 2: Network view on the renin-angiotensin system (*hsa04614*). Nodes are encoded according to category assignment as triangles (multiple sources), octagons (pathophysiology-omics), diamonds (patents), and circles (not identified as affected in any of the given data sets). Edges represent relations as provided from omicsNET. Underlined gene symbols were assigned to this pathway based on the extension procedure applied.


A
CC-lit
CC-pat
4
730
2
59
25
876
24
0
80
14
96
0
10
5
10
PP
CC-trials
B
Proteomics
Metabolomics
25
87
0
2
1
695
2
0
30
0
4
0
0
1
4
Transcriptomics
SNP



CCK
ACE2
AGT
ACE
CMA1
REN
KLK7
ENPEP
DPEP
LNPEP
XPNPEP2