Molecular models of the cardiorenal syndrome

Andreas Heinzel[1], Raul Fechete[1], Irmgard Mühlberger[1], Paul Perco[1], Bernd Mayer[1,2], Arno Lukas[1]

[1]emergentec biodevelopment GmbH

Gersthofer Strasse 29-31, 1180 Vienna, Austria

[2]Institute for Theoretical Chemistry, University of Vienna

Währingerstrasse 17, 1090 Vienna, Austria

Correspondence:

Dr. Arno Lukas, emergentec biodevelopment GmbH, Gersthofer Strasse 29-31, 1180 Vienna, Austria

E-mail: arno.lukas@emergentec.com

Fax: +43 1 4034966-19

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may leadto differences between this version and the Version of Record. Please cite this articleas doi: 10.1002/elps.201200642

Received: November 27, 2012; Revised: February 8, 2013; Accepted: February 13, 2013

List of Abbreviations:

ATS, atherosclerosis

CAD, coronary artery disease

CKD, chronic kidney disease

CVD, cardiovascular disease

CRS, cardiorenal syndrome

DN, diabetic nephropathy

ESRD, end stage renal disease

GEO, Gene Expression Omnibus

MACE, major adverse cardiac event

MeSH, Medical Subject Headings

NCBI, National Center for Biotechnology Information

Keywords: systems biology, cross-omics, integration, process, biomarker

Total number of words (including figure and table legends): 5784

Number of display elements: 5

Abstract

Molecular profiling techniques have provided extensive sets of molecular features characterizing clinical phenotypes, but further extrapolation to mechanistic molecular models of disease pathophysiology faces major challenges. Here we describe a computational procedure for delineating molecular disease models utilizing omics profiles, and exemplify the methodology on aspects of the cardiorenal syndrome describing the clinical association of declining kidney function and increased cardiovascular event rates. Individual molecular features as well as selected molecular processes were identified as linking cardiovascular and renal pathology as a combination of cross-organ mediators and common pathophysiology. The molecular characterization of the disease presents as a set of molecular processes together with their interactions, composing a molecular disease model of the cardiorenal syndrome.

Integrating omics profiles describing aspects of cardiovascular disease and respective profiles for advanced chronic kidney disease on molecular interaction networks, computation of disease term-specific subgraphs, and complemented by subgraph segmentation allowed delineation of disease term-specific molecular models, at their intersection providing contributors to cardiorenal pathology.

Building such molecular disease models allows in a generic way to integrate multi-Omics sources for generating comprehensive sets of molecular processes, on such basis providing rationale for biomarker panel selection for further characterizing clinical phenotypes.

1 Introduction

Within the last decade Omics techniques, spanning from genomic sequencing to metabolite profiling, have significantly enlarged our understanding of molecular processes characterizing pathophysiology of human diseases. Such understanding is deemed vital for identification of biomarkers as well as drug targets being mechanistically involved in a specific pathophysiology and subsequent clinical presentation. For obtaining such causalities with descriptive Omics signatures as input (being sets of molecular features identified as associated with disease), delineation of processes and pathways has been emerging as central concept for overcoming the gap between disease association and pathophysiological causality. Such analysis aims at integrating the molecular data space describing a specific clinical phenotype together with background knowledge regarding involvement of such disease associated features in intra- as well as intercellular processes and pathways. This route promises identification of mechanisms associated with disease, in contrast to sequentially analyzing individual features being in a first place associated on a statistical level with a specific clinical presentation. Systems Biology is mainly driving such concepts and methods [1], in the realm of biomedical research comprising efforts in Systems- and Pathway Medicine [2].

Deciphering processes and pathways as relevant regarding a clinical presentation, however, sees numerous challenges, one fundamental issue being the clinical cataloguing of phenotypes in terminology of disease/syndrome as well as signs/symptoms (as e.g. followed in NCBI MeSH). Specifically, processes and pathways are used for describing peculiarities of a perturbed molecular state, which on a clinical level presents with certain characteristics. Clinical terminology on the other hand describes a phenotype which may encompass indeed a specific molecular process, but frequently is comprised of a heterogeneous molecular background, in turn needing multiple processes and pathways for correctly describing such phenotype. Specifically for age-associated, chronic diseases this fact becomes evident, where e.g. in MeSH the disease term ‘cardiovascular disease’ holds 405 specific terms, being a mixture of event-centric nomenclature (as ‘stroke’), of symptoms (as ‘hypertension’) and of

pathology (as 'cardiomyopathy'). On top, again specifically in age-associated disorders, clinical presentation is frequently a combination of morbidities, further hampering assignment of sensitive but still specific molecular processes and pathways. For certain disease terms this multifactorial nature becomes already evident on the terminology level, with diabetic nephropathy (implications of diabetes mellitus on kidney function) or the cardiorenal syndrome (CRS) as examples.

The CRS describes a combined clinical presentation where either the cardiovascular system or kidney damage is implicating deleterious processes in the respective other organ, resulting in an aggregate clinical presentation of coupled cardiovascular system (specifically heart) and renal disease. For providing a more detailed clinical categorization, and taking the original organ damage, acute or chronic presentation, and systemic disease as diabetes into account, Ronco et al. categorized the CRS further into five subtypes [3, 4]. In general, rate of death for any cause increases sharply with declining kidney function (being at about 20% annually in end stage renal disease (ESRD) [5]), with major contributions from cardiovascular mortality [6].

Conceptually, two general molecular mechanisms may be relevant for describing the pathophysiology of the CRS: One route may be organ cross-talk, i.e. mediators released by one organ trigger pathological processes (as fibrosis, atherosclerosis) at the other. A second concept is joint pathophysiology as e.g. endothelial dysfunction, oxidative stress, or vascular inflammation, being causative for joint function decline of kidney and cardiovascular system [7, 8]. Integrative analysis of organ-specific transcriptomics profiles together with molecular features identified as being CRS-associated using scientific literature mining together with omics profiling data allowed addressing both aspects [9, 10]. Such analyses provided results on well established key processes as alterations in the activity of the renin-angiotensin system and mediators as the natriuretic peptide receptor and ligands. Next to hemodynamics an accumulation of CRS -associated features was identified in the endothelin, VEGF signaling, PPAR signaling, coagulation, inflammation, and focal adhesion pathways. In summary, such integrated view pinpointed mechanisms associated with dyslipidemia, hemodynamic regulation, and inflammation on the interface between the cardiovascular and the renal system as being involved. Based on the given result status both mechanisms, i.e.

mediators together with common pathophysiology, appear to contribute to the clinical presentation of the CRS in general, presumably with specific relative contributions in the five CRS subtypes.

Although adding to our knowledge regarding the relevance of specific features and involvement of selected pathways mostly in isolation, integrated molecular models explicitly linking mediators, processes and pathways are not yet available for characterizing the CRS. As recently demonstrated for diabetic nephropathy [11], and laid out conceptually in Heinzel et al. [12], molecular models of disease terms, composed of a set of disease-associated processes and their interrelation, promise improved interpretation of molecular mechanisms regarding their causality as well as relative contribution to a clinical presentation. Further, such molecular models may improve biomarker candidate selection, allowing monitoring the apparently complex, multi-process pathophysiology of the CRS, in turn also allowing hypothesis generation regarding novel therapy targets.

We here present a generic strategy for delineating molecular disease models resting on molecular profiling data, specifically aiming at analyzing the CRS characterized by primary chronic kidney disease and cardiac dysfunction. We combine Omics profiling data available in the public domain for established chronic kidney disease (CKD) with specific focus on diabetic nephropathy and cardiovascular disease (CVD) phenotypes, namely coronary artery disease and atherosclerosis, for delineating disease term-specific molecular models, followed by generation of an interference molecular model for delineating molecular processes characterizing the reno-cardiac axis.

2 Materials and methods

2.1 Molecular data sets

For describing CVD and CKD on a molecular feature level, public domain data from transcriptomics studies as well as from scientific literature were retrieved. Transcriptomics data sets were selected following a literature search in NCBI PubMed

(http://www.ncbi.nlm.nih.gov/pubmed/) with specific CVD and CKD disease MeSH terms as major topics (Figure 1).

insert Figure 1 here

For reducing disease term heterogeneity, CKD data retrieval was restricted to diabetic nephropathy, being the most prevalent cause of ESRD. For CVD two disease terms of relevance regarding major adverse cardiac events (MACE) in established CKD were selected, namely coronary artery disease (CAD) and atherosclerosis (ATS). Publications identified for these terms were further restricted to transcriptomics studies (microarray analysis, oligonucleotide array sequence analysis, gene expression profiling) using human samples. Manual curation restricted the eligible data sets further for only including studies using kidney biopsies (all compartments as well as microdissected material) on the CKD, or arteries on the CVD side in case/control study design. This procedure yielded three eligible transcriptomics studies on CKD, and four on CVD, respectively.

For CKD studies performed by Woroniecka et al. [13], Berthier et al. [14] and Cohen et al. [15], the set of differentially regulated genes comparing CKD and control were retrieved as reported in the original publications, equivalently handling CVD studies provided by Volger et al. [16], Archaki et al. [17] and Cagnin et al. [18]. The CVD study of Haegg et al. [19] did not provide such a result set and therefore required statistical analysis of respective raw data for obtaining a result list of differentially regulated features. For this the study-specific Affymetrix Human Genome U133 Plus 2.0 Array CEL files were retrieved from GEO (http://www.ncbi.nlm.nih.gov/geo/, data identifier GSE40231) and analyzed using CARMAweb [20] (applying background correction, summarization of probe set values, and MAS5 normalization), followed by significance analysis of microarrays (SAM) [21] for identifying differentially expressed genes (false discovery rate of ≤ 5%, minimum fold change of 2.0). However, eventual presence of the CRS cannot be excluded for the selected datasets, as patients with established diabetic nephropathy show substantially increased risk for major cardiovascular events [5], together with e.g. hypertensive nephrosclerosis seen for patients with cardiovascular complications [22].

Complementary, a literature mining approach utilizing Pubmed MeSH annotation and information provided by gene2pubmed [23] was performed. For CKD molecular feature retrieval was performed using the term “kidney failure, chronic”. CVD molecular feature retrieval used the MeSH terms (again as major terms) “atherosclerosis” and “coronary artery disease”. Finally, all molecular features retrieved from either transcriptomics or scientific literature were mapped to Ensembl gene identifiers for allowing subsequent integrative analysis (Table 1).

insert Table 1 here

2.2 Interaction network and molecular disease models

As basis for delineating molecular disease models a human protein coding gene interaction network was used [11, 24] representing 14,021 protein coding genes (out of in total 20,038 protein coding genes as represented in Ensembl [25]) as network nodes and holding 694,045 relations as edges between nodes. Such relations included consolidated protein interaction information from IntAct [26], BioGRID [27] and Reactome [28] (database versions as of August 2012), totaling in 176,977 interactions. Further relations were inferred by integrating pathway assignment from Reactome and PANTHER [29], ontology process and function from the Gene Ontology [30], and protein domains from InterPro [31]. Each relation was assigned with a weight, with a maximum number of 1.0 for relations holding explicit experimental background, and weights from the interval [0.0, 1.0] for inferred relations for reflecting the evidence of such relations. For the specific analysis an edge weight cutoff of 0.74 was used, providing us with an optimal recovery of experimentally known interactions, and totaling in a number of edges as speculated for the human protein interactome [32].

Consolidated molecular feature sets on CVD and CKD (Table 1) were mapped on this hybrid relations network (with 2,948 protein coding genes assigned to CKD also holding a node in the network, and correspondingly 3,556 assigned network nodes for CVD), subsequently removing all network nodes (protein coding genes) not being associated with either CVD or CKD. Result of this procedure is disease term-specific subgraphs for CKD and CVD, respectively. Segmentation of such subgraphs was performed by applying the MCODE algorithm [33] using default settings. The algorithm extracts highly

interconnected regions from the graph by weighting all vertices according to the density of their local neighborhood and subsequent outward traversal starting from the densest region of the weighted graph. MCODE was originally designed to identify molecular complexes in protein-protein interaction networks. However, in a hybrid relations network incorporating protein binding as well as functional and procedural dependencies such identified regions may expand beyond mere complexes to representing molecular processes.

Aggregate relations between such segments (processes) were computed as ratio of the number of relations effectively found between members of segments and the theoretical number of such relations (being n x m, with n the number of molecular features in one segment, and m the respective number in the other segment). The set of segments together with inter-segment relations composed the molecular disease models. Visualization of subgraphs and molecular disease models were done in gephi [34].

For delineating an interference molecular model on the basis of the individual molecular disease models for CKD and CVD, the significance of enrichment of features belonging to a specific segment of one molecular disease model was evaluated in each individual segment of the other molecular disease model. A Fisher's exact test including correction for multiple testing (FDR ≤ 5%) was applied. For such sets of shared features KEGG pathway enrichment analysis [35] was done utilizing DAVID [36].

3 Results

According to Table 1 3,188 protein coding genes could be assigned to CKD, and 3,770 to CVD, with 1,115 molecular features being shared in both data sets. For CKD 144 features were identified in both transcriptomics and literature, the respective number for CVD is 260. The majority of features were retrieved from transcriptomics studies, however, seeing substantial variation in number of features identified in individual studies, as well as minor feature overlap between studies. This appears as common finding in transcriptomics profile meta-analysis [24, 37, 38], next to heterogeneity in the

experimental profiling procedures also resting on specific study design and details of inclusion criteria. Integrating the various sources on the one hand promises a more complete characterization of the disease terms (presumably still seeing false negatives due to incomplete representation of the relevant transcriptome on platforms commonly used, but also due to frequently small cohorts analyzed), but also coming in hand with an increased false positive rate. Previous studies, however, indicated that comparing profiles on an individual feature level overestimates heterogeneity, showing increased homogeneity when consolidating profiles on pathways and processes and then comparing feature sets on this aggregate level [24].

Mapping the molecular feature set for CKD and CVD on the reference relations network allowed assignment of 92% of given CKD and of 94% of given CVD features (with the others not being included in the reference network). After removing all network nodes not being part of the consolidated CKD or CVD feature lists the disease term-specific subgraphs resulted (Figure 2).

insert Figure 2 here

Of the in total 2,948 CKD features being present on the reference network 555 features did not show a single relation (network edge) to another CKD feature (on the level of the relations network used) of the given set. Such features were identified as relevant in a specific study according to Table 1, but on a process level (for which the reference network is deemed a proxy) no link to any other relevant feature became apparent. The respective number for the CVD set (with a total of 3,556 features being on the network) was 680. Such features may be attributed as false positive features from multi-source integration, and were removed from further considerations, leaving 2,393 features for CKD and 2,876 features for CVD.

The disease term-specific subgraphs as presented in Figure 2 were forwarded to network segmentation utilizing the MCODE algorithm. This procedure provided 44 segments (sets of nodes) for CKD holding in total 933 features, and 50 segments for CVD holding in total 1,109 features. This process resulted in a further substantial decrease of disease term-specific features for further analysis, potentially removing relevant features regarding explanation of molecular pathophysiology. On the other hand, this

procedure extracted features which next to their disease association as such also exhibited a high degree of interactions to other members of such segment (all being disease associated features), and based on the optimization criteria of the MCODE algorithm less such connectivity to members of a different segment. In the following such segment is interpreted as disease term-specific molecular process, and computing the aggregate relation of features between such segments provided a molecular disease model for CKD and CVD (Figure 3).

insert Figure 3 here

The smallest number of features in a segment was found to be 3, with a maximum number of features in a segment identified as 156 for CKD, and 201 for CVD respectively. The 75% quantile was 15.5 for CKD and 15.0 for CVD, respectively. For obtaining a molecular model specifically addressing processes of relevance for the CRS the overlap of segments when comparing CKD and CVD model was computed, resulting in 2,200 such comparisons (with 44 segments on the CKD, and 50 segments on the CVD side). 61 cross-disease segment pairs were found to share at least one molecular feature, while all others did not overlap on a single molecular feature level (hence being CKD or CVD specific). The CKD model contributed with 29 unique segments of which 15 showed an overlap with only a single CVD segment. 14 CKD segments showed overlaps with two to six different CVD segments. The CVD model contributed with 30 unique segments, of which 16 showed an overlap with only a single CKD segment each. 14 CVD segments provided an overlap with two to seven different CKD segments each. Two segments showed an overlap with two other segments for the CVD and CKD model, respectively. When testing for significance of such cross-segment/cross-disease model enrichment 21 segment pairs out of the 61 sharing at least one feature remained as significant for further analysis. For these the CKD and CVD model each contributed with 18 unique segments, of which 15 each were overlapping with only one segment of the other disease model. Three were overlapping with 2 segments of the other disease model each. Minimum number of shared features varied between 2 and 41 (75% quantile of 16) (Figure 4).

insert Figure 4 here

As an attempt to further annotate these junctions between CKD and CVD the overlap of each of the 21 shared feature sets themselves were forwarded to enrichment analysis utilizing KEGG pathways as reference. For three feature set pairs significant enrichment of KEGG pathways could be identified. The first was found to involve the largest segments of both models, namely CKD-U2 (156 features) and CVD-U1 (201 features). The 41 overlapping features showed significant enrichment regarding the KEGG pathways chemokine signaling ($p=4.33 * 10^{-8}$) and cytokine-cytokine receptor interaction ($p=3.75 * 10^{-4}$). The segment CVD-U1 is also part of a second overlap with segment CKD-U1 (96 features) holding 35 joint features. This feature set was found as significantly associated with focal adhesion ($p=1.02 * 10^{-3}$), adherence junction ($p=6.06 * 10^{-4}$) and endocytosis ($p=1.33 * 10^{-6}$). As third enriched segment overlap CKD-U6 (115 features) with CVD-U8 (32 features) was identified, holding 16 joint features showing significant association with retinol metabolism ($p=5.37 * 10^{-3}$), drug metabolism ($p=9.37 * 10^{-3}$), arachidonic acid metabolism ($p=6.22 * 10^{-3}$), linoleic acid metabolism ($p=3.61 * 10^{-4}$) and metabolism of xenobiotics by cytochrome P450 ($p=8.21 * 10^{-3}$).

4 Discussion

Omics profiling in biomedical research has substantially contributed to knowledge on individual molecular features and in part processes and pathways associated with clinical phenotypes. However, traversing such data into hypotheses on novel biomarkers and therapy targets and subsequent clinical use has been less effective. We in this work aim at specifically addressing two main issues relevant in this context, being i) methodology for turning descriptive feature lists (be it from transcriptomics, proteomics, or other profiling sources or literature mining) deemed characteristic for a clinical phenotype in a molecular disease model, and ii) linking disease terms and underlying molecular pathophysiology as such, here exemplified on aspects of the cardiorenal syndrome.

Improved understanding of the cardiorenal syndrome is of significant relevance, specifically in late stages of chronic kidney disease seeing significant cardiovascular event rates. Following the concepts of mediators and common pathophysiology,

potentially also their coexistence, may allow identifying biomarkers for MACE risk assessment, and from there potentially novel therapy targets aimed at modulating deleterious cross-talk mechanisms. In previous work we identified individual components contributing to mechanisms of relevance in the cardiorenal setting, elements again becoming evident when interfering molecular models of CKD and CVD disease terms. In total 21 segments of the CKD and CVD molecular model showed overlap, three of them allowing interpretation in the context of KEGG pathway enrichment information. One aspect identified is focal adhesion, adherens junction and the endocytosis pathway. All three pathways have been implicated in the process of cell migration [39] being of relevance in CVD. In addition, Silverstein et al. reported in the very context that expression of adhesion molecules stimulated by inflammation of renal tissue contributes to vascular stiffening [40]. For the second overlap allowing annotation in KEGG the retinol metabolism, arachidonic acid metabolism, linoleic acid metabolism, drug metabolism and metabolism of xenobiotics by cytochrome P450 were identified, with an association of plasma retinol concentration and CVD mortality previously reported in [41]. Identification of the fatty acid metabolism pathways can be explained by inflammation due to renal dysfunction causing changes in the blood lipid composition and as such contributing to the development of CVD [42]. Finally, identification of chemokine signaling as well as the cytokine-cytokine receptor interactions reflect the involvement of inflammatory processes [43], also identified as relevant when specifically analyzing diabetic nephropathy [37].

Besides supportive data from conventional pathway enrichment analysis at the interface of CKD and CVD molecular models, further evidence also on a single feature level supports the presented concept. From the in total 223 molecular entities taking part in the interference model, 40 were explicitly discussed in previous work on the CRS [9]. Among them twelve blood pressure regulating features play a role (as ACE, AGT) [44], naturally being of relevance for the cardiorenal axis [45]. In addition another vasoconstrictor, namely UTS2, was part of the interference feature set. Zhu et al. observed that significant changes in expression levels of UTS2 can be detected in cardiovascular and renal disease patients [46].

When analyzing direct interference between molecular disease models not only shared features, but also linking features may be analyzed. Such linkers are molecular entities which, in terms of betweenness centrality, are not important on the level of a segment but become relevant when the segment are interfaced via shared features. Using such method as outlined in [47] 508 such linkers for interfering CKD and CVD units could be identified. Forwarding this set of linker features to KEGG pathway enrichment analysis identified four of the above discussed pathways, namely chemokine signaling ($p=5.32 * 10^{-10}$), focal adhesion ($p=5.73 * 10^{-10}$), adherens junction ($p=1.09 * 10^{-3}$) and endocytosis ($p=2.88 * 10^{-2}$), together with a number of further pathways in part already discussed in the renal or cardial context as PI3K-Akt signaling ($p=8.06 * 10^{-9}$), HIF-1 signaling ($p=4.12 * 10^{-7}$) and VEGF signaling ($p=6.41 * 10^{-5}$).

Certainly, the concept of molecular models as discussed in this work sees numerous challenges. Availability of thorough profiling data of clinical phenotypes is required to ensure that subsequently extracted disease networks capture all relevant molecular aspects including the cross-talk with other diseases under investigation. Specifically for the latter a method for identifying pleiotropic molecular entities [48] could be used for including features eventually missed during disease graph extraction. Further improvements of coverage and evidence of reference interaction networks, improved segmentation procedures specifically regarding the topology of biological networks [49], and the assessment of indirect effects among molecular entities of different diseases [50] are of major importance to the concept. On the other hand, the concept is versatile, allowing an integration of different Omics profiles for truly covering genetic predisposition as well as environmental impact in focus of risk for developing disease and disease progression [11, 37]. Ultimately, the true strength of molecular models may be the more complete representation of pathophysiology for given disease terminology [12] as an assembly of causative processes which, on a population level, show patient-specific contributions and relevance. By including the total set of processes, and consequently biomarker panels instead of single markers, present limitations in sensitivity and specificity in diagnosis and prognosis in complex disorders as the cardiorenal syndrome may be overcome.

References:

[1] Molina, F., Dehmer, M., Perco, P., Graber, A., Girolami, M., Spasovski, G., Schanstra, J. P., Vlahou, A., *Nephrol Dial Transplant.* 2010, *25*, 1015-1018.

[2] Auffray C., Chen Z., Hood L., *Genome Med.* 2009, *1*, 2.

[3] Ronco, C., Maisel, A., *Congest Heart Fail.* 2010, *16 Suppl 1*, 1-4.

[4] Ronco, C., Haapio, M., House, A. A., Anavekar, N., Bellomo, R., *J Am Coll Cardiol.* 2008, *52*, 1527-1539.

[5] Adler, A. I., Stevens, R. J., Manley, S. E., Bilous, R. W., Cull, C. A., Holman, R. R., *Kidney Int.* 2003, *63*, 225-32.

[6] Sarnak, M. J., Levey, A. S., Schoolwerth, A. C., Coresh, J., Culleton, B., Hamm, L. L., McCullough, P. A., Kasiske, B. L., Kelepouris, E., Klag, M. J., Parfrey, P., Pfeffer, M., Raij, L., Spinosa, D. J., Wilson, P. W., *Circulation*. 2003, *108,* 2154-2169.

[7] Shestakova, M. V., Jarek-Martynowa, I. R., Ivanishina, N. S., Kuharenko, S. S., Yadrihinskaya, M. N., Aleksandrov, A. A., Dedov, I. I., *Diabetes Res Clin Pract.* 2005, *68 Suppl1*, 65-72.

[8] Yagi, S., Aihara, K., Ikeda, Y., Akaike, M., Sata, M., Matsumoto, T., *Int J Vasc Med.* 2012, 162545.

[9] Mühlberger, I., Mönks, K., Bernthaler, A., Jandrasits, C., Mayer, B., Mayer, G., Oberbauer, R., Perco, P., *Int J Nephrol.* 2011, 809378.

[10] Mühlberger, I., Mönks, K., Fechete, R., Mayer, G., Oberbauer, R., Mayer, B., Perco, P., *OMICS.* 2012, *16*, 105-112.

[11] Mayer, P., Mayer, B., Mayer, G., *Nephrol Dial Transplant*. 2012, *27*, 3995-4002.

[12] Heinzel, A., Fechete, R., Söllner, J., Perco, P., Heinze, G., Oberbauer, R., Mayer, G., Lukas, A., Mayer, B., *Int J of Sys Bio and Biomed Tech.* 2012, *1*, 11-25.

[13] Woroniecka, K. I., Park, A. S., Mohtat, D., Thomas, D. B., Pullman, J. M., Susztak, K., *Diabetes. 2011*, *60*, 2354-2369.

[14] Berthier, C. C., Zhang, H., Schin, M., Henger, A., Nelson, R. G., Yee, B., Boucherot, A., Neusser, M. A., Cohen, C. D., Carter-Su, C., Argetsinger, L. S., Rastaldi, M. P., Brosius, F. C., Kretzler, M., *Diabetes.* 2009, *58*, 469-477.

[15] Cohen, C. D., Lindenmeyer, M. T., Eichinger, F., Hahn, A., Seifert, M., Moll, A. G., Schmid, H., Kiss, E., Gröne, E., Gröne, H. J., Kretzler, M., Werner, T., Nelson, P. J., *PLoS One.* 2008, *3*, e2937.

[16] Volger, O. L., Fledderus, J. O., Kisters, N., Fontijn, R. D., Moerland, P. D., Kuiper, J., van Berkel, T. J., Bijnens, A. P., Daemen, M. J., Pannekoek, H., Horrevoets, A. J., *Am J Pathol.* 2007, *171*, 326-337.

[17] Archacki, S. R., Angheloiu, G., Tian, X. L., Tan, F. L., DiPaola, N., Shen, G. Q., Moravec, C., Ellis, S., Topol, E. J., Wang, Q., *Physiol Genomics.* 2003, *15*, 65-74.

[18] Cagnin, S., Biscuola, M., Patuzzo, C., Trabetti, E., Pasquali, A., Laveder, P., Faggian, G., Iafrancesco, M., Mazzucco, A., Pignatti, P. F., Lanfranchi, G., *BMC Genomics.* 2009, *10*, 13.

[19] Haegg, S., Skogsberg, J., Lundström, J., Noori, P., Nilsson, R., Zhong, H., Maleki, S., Shang, M. M., Brinne, B., Bradshaw, M., Bajic, V. B., Samnegård, A., Silveira, A., Kaplan, L. M., Gigante, B., Leander, K., de Faire, U., Rosfors, S., Lockowandt, U., Liska, J., Konrad, P., Takolander, R., Franco-Cereceda, A., Schadt, E. E., Ivert, T., Hamsten, A., Tegnér, J., Björkegren, J., *PLoS Genet.* 2009, *5*, e1000754.

[20] Rainer, J., Sanchez-Cabo, F., Stocker, G., Sturn, A., Trajanoski, Z., *Nucleic Acids Res.* 2006, *34*, W498-W503.

[21] Tusher, V. G., Tibshirani, R., Chu, G., *Proc Natl Acad Sci U S A.* 2001, *98*, 5116-5121.

[22] Luft, F. C., Curr Opin Nephrol Hypertens. 2004, 13(2), 147-54.

[23] Maglott, D., Ostell, J., Pruitt, K. D., Tatusova, T., *Nucleic Acids Res.* 2007, *35*, D26-31.

[24] Bernthaler, A., Mühlberger, I., Fechete, R., Perco, P., Lukas, A., Mayer, B., *Mol Biosyst.* 2009, *5*, 1720-31.

[25] Flick, P., Amode, M. R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gil, L., Gordon, L., Hendrix, M., Hourlier, T., Johnson, N., Kähäri, A. K., Keefe, D., Keenan, S., Kinsella, R., Komorowska, M., Koscielny, G., Kulesha, E., Larsson, P., Longden, I., McLaren, W., Muffato, M., Overduin, B., Pignatelli, M., Pritchard, B., Riat, H. S., Ritchie, G. R., Ruffier, M., Schuster, M., Sobral, D., Tang, Y. A., Taylor, K., Trevanion, S., Vandrovcova, J., White, S., Wilson, M., Wilder, S. P., Aken, B. L., Birney, E., Cunningham, F., Dunham, I., Durbin, R., Fernández-Suarez, X. M., Harrow, J., Herrero, J., Hubbard, T. J., Parker, A., Proctor, G., Spudich, G., Vogel, J., Yates, A., Zadissa, A., Searle, S. M., *Nucleic Acids Res.* 2012, *40*, D84-90.

[26] Kerrien, S., Aranda, B., Breuza, L., Bridge, A., Broackes-Carter, F., Chen, C., Duesbury, M., Dumousseau, M., Feuermann, M., Hinz, U., Jandrasits, C., Jimenez, R. C., Khadake, J., Mahadevan, U., Masson, P., Pedruzzi, I., Pfeiffenberger, E., Porras, P., Raghunath, A., Roechert, B., Orchard, S., Hermjakob, H., *Nucleic Acids Res.* 2012, *40*, D841-846.

[27] Stark, C., Breitkreutz, B. J., Reguly, T., Boucher, L., Breitkreutz, A., Tyers, M., *Nucleic Acids Res.* 2006, *34*, D535-539.

[28] Matthews, L., Gopinath, G., Gillespie, M., Caudy, M., Croft, D., de Bono, B., Garapati, P., Hemish, J., Hermjakob, H., Jassal, B., Kanapin, A., Lewis, S., Mahajan, S., May, B., Schmidt, E., Vastrik, I., Wu, G., Birney, E., Stein, L., D'Eustachio, P., *Nucleic Acids Res.* 2009, *37*, D619-622.

[29] Thomas, P. D., Campbell, M. J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., Diemer, K., Muruganujan, A., Narechania, A., *Genome Res.* 2003, *13*, 2129-2141.

[30] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., Sherlock, G., *Nat Genet.* 2000, *25*, 25-29.

[31] Hunter, S., Jones, P., Mitchell, A., Apweiler, R., Attwood, T. K., Bateman, A., Bernard, T., Binns, D., Bork, P., Burge, S., de Castro, E., Coggill, P., Corbett, M., Das, U., Daugherty, L., Duquenne, L., Finn, R. D., Fraser, M., Gough, J., Haft, D., Hulo, N., Kahn, D., Kelly, E., Letunic, I., Lonsdale, D., Lopez, R., Madera, M., Maslen, J., McAnulla, C., McDowall, J., McMenamin, C., Mi, H., Mutowo-Muellenet, P., Mulder, N., Natale, D., Orengo, C., Pesseat, S., Punta, M., Quinn, A. F., Rivoire, C., Sangrador-Vegas, A., Selengut, J. D., Sigrist, C. J., Scheremetjew, M., Tate, J., Thimmajanarthanan, M., Thomas, P. D., Wu, C. H., Yeats, C., Yong, S. Y., *Nucleic Acids Res.* 2012, *40*, D306-312.

[32] Stumpf, M. P., Thorne, T., de Silva, E., Stewart, R., An, H. J., Lappe, M., Wiuf, C., Proc Natl Acad Sci U S A. 2008, 105(19), 6959-64.

[33] Bader, G. D., Hogue, C. W. V., *BMC Bioinformatics.* 2003, *4*, 2.

[34] Bastian, M., Heymann, S., Jacomy, M., *Inter AAAI Conf on Weblogs and Social Media.* 2009.

[35] Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., Tanabe, M., *Nucleic Acids Res.* 2012, *40*, D109-114.

[36] Huang, D. W., Sherman, B. T., Lempicki, R. A., *Nature Protoc.* 2009, *4*, 44-57.

[37] Fechete, R., Heinzel, A., Perco, P., Mönks, K., Söllner, J., Stelzer, G., Eder, S., Lancet, D., Oberbauer, R., Mayer, G., Mayer, B., *Proteomics Clin Appl.* 2011, *5*, 354-366.

[38] Rapberger, R., Perco, P., Sax, C., Pangerl, T., Siehs, C., Pils, D., Bernthaler, A., Lukas, A., Mayer, B., Krainer, M., *BMC Syst Biol.* 2008, *2*, 2.

[39] Kawauchi, T., *Int J Mol Sci.* 2012, *13*, 4564-4590.

[40] Silverstein, D. M., *Pediatr Nephrol.* 2009, *24*, 1445-1452.

[41] Brazionis, L., Walker, K. Z., Itsiopoulos, C., O'Dea, K., *Nutr Metab Cardiovasc Dis.* 2012, *22*, 914-920.

[42] Schiffrin, E. L., Lipman, M. L., Mann, J. F., *Circulation* 2007, *116*, 85-97.

[43] Colombo, P. C., Ganda, A., Lin, J., Onat, D., Harxhi, A., Iyasere, J. E., Uriel, N., Cotter, G., *Heart Failure Rev.* 2012, *17*, 177-190.

[44] Judson, W. E., Helmer, O. M., C*irculation.* 1971, *44*, 245-253.

[45] Longhini, C., Molino, C., Fabbian, F., *Clin Exp Nephrol.* 2010, *14*, 12-21.

[46] Zhu, Y. C., Zhu, Y. Z., Moore, P. K., *Br J Pharmacol.* 2006, *148*, 884-901.

[47] Vaggi, F., Dodgson, J., Bajpai, A., Chessel, A., Jordán, F., Sato, M., Carazo-Salas, R. E., Csikász-Nagy, A., PLoS Comput Biol. 2012, 8(10), e1002732.

[48] Nguyen, T. P., Liu, W. C., Jordán, F., BMC Syst Biol. 2011, 5:179.

[49] Wang J., Li M., Deng Y., Pan Y., *BMC Genomics. 2010*, *11 Suppl 3*, 10.

[50] Nguyen, T. P., Jordán, F., BMC Syst Biol. 2010, 4:103.

Acknowledgements

Financial support for this study was obtained from Fresenius Medical Care Austria GmbH.

Conflict of interest statement

The authors declare no conflict of interest.

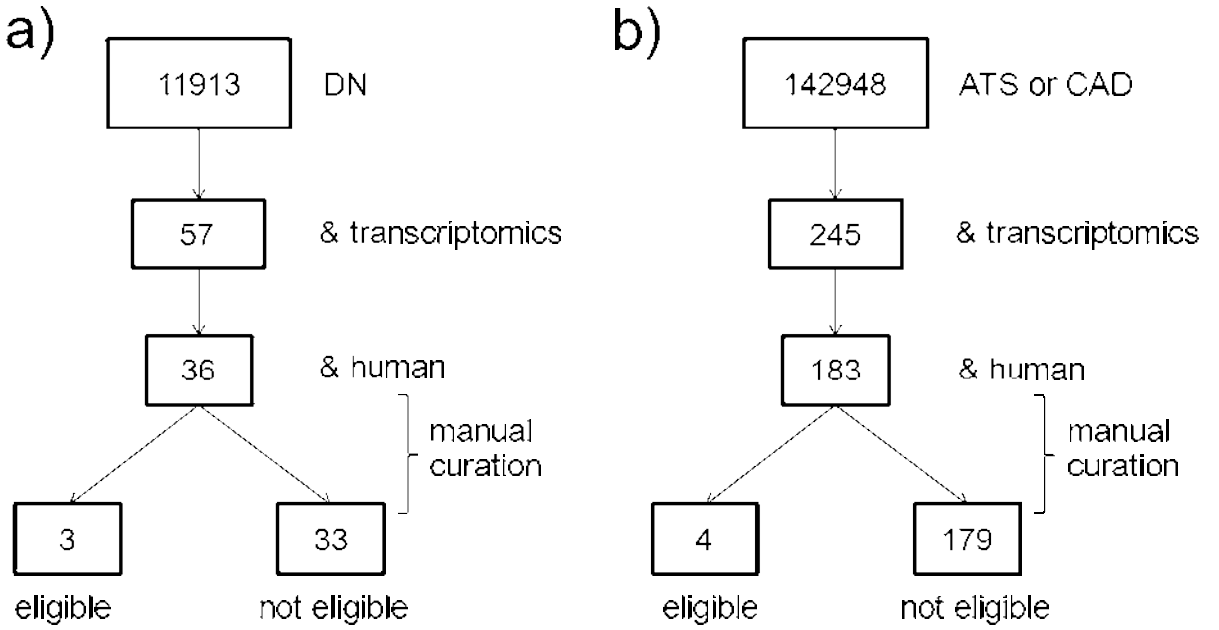


Figure 1: Transcriptomics studies retrieval process for a) CKD with focus on DN and b) CVD focusing on ATS and CAD. Provided is the number of studies as identified in Pubmed being eligible at a specific step of curation.

data set	disease term	tissue	# of features CKD	# of features CVD	# of features overlap
Woroniecka et al.	DN	whole kidney	2779	-	-
Berthier et al.	DN	tubulointerstitium	7	-	-
Cohen et al.	DN	tubulointerstitium	70	-	-
Haegg et al.	CAD	aorta / wall		625	-
Volger et al.	ATS	large arteries / endothelium	-	1113	-
Archaki et al.	CAD	coronary arteries/wall	-	90	-
Cagnin et al.	ATS	coronary arteries/wall	-	1070	-
transcriptomics feature sets, total			2815	2698	642
literature feature sets, total			517	1332	286
consolidated feature sets, total			3188	3770	1115

Table 1: Data references for transcriptomics and literature data sets. Provided is the specific disease term, tissue used for transcriptomics, number of unique features per study, and feature overlap.

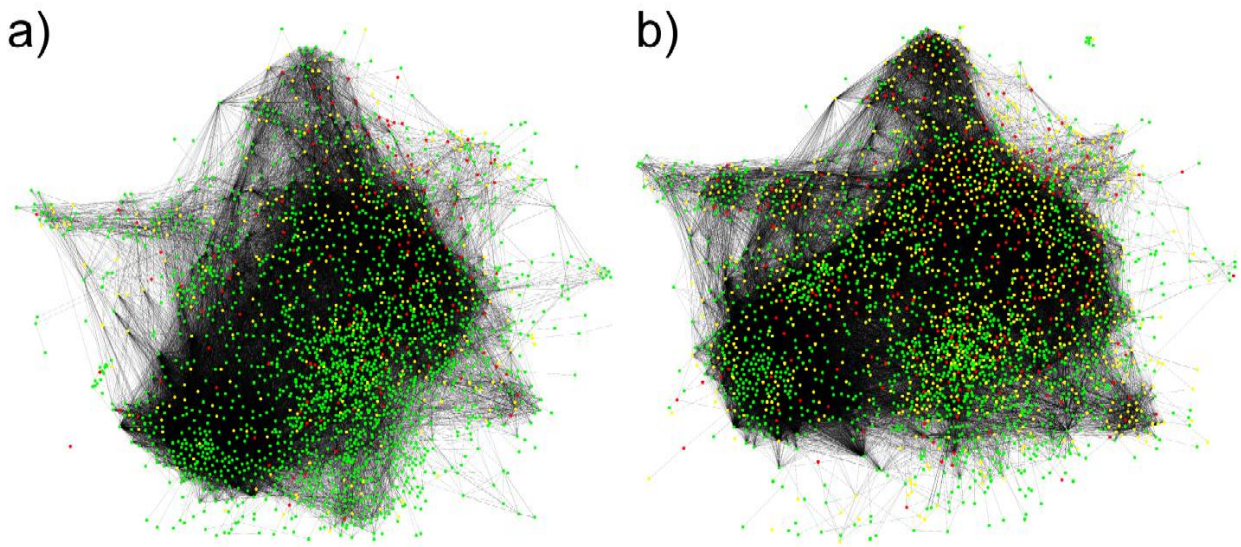


Figure 2: Disease term-specific subgraphs for a) CKD and b) CVD. Displayed are nodes (relevant protein coding genes) as identified and mapped on the reference network, and their relations (edges) according to the specific reference network used. Nodes provided in red were identified via transcriptomics and literature extraction, nodes in green result from transcriptomics only, nodes in yellow from literature extraction only.

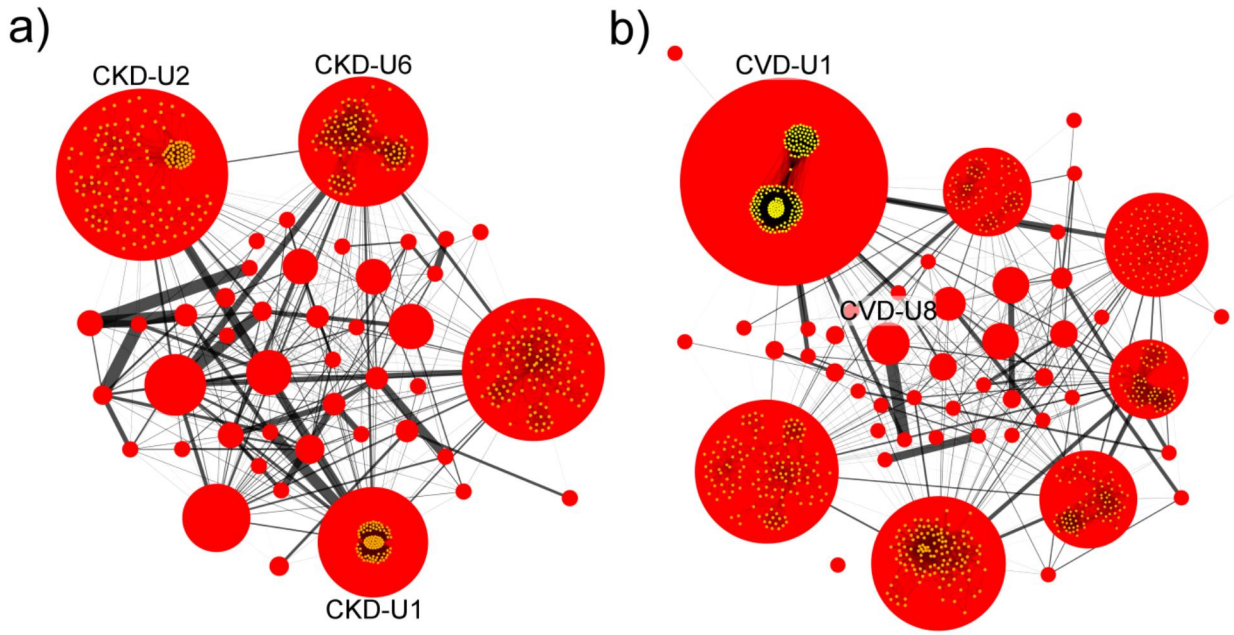
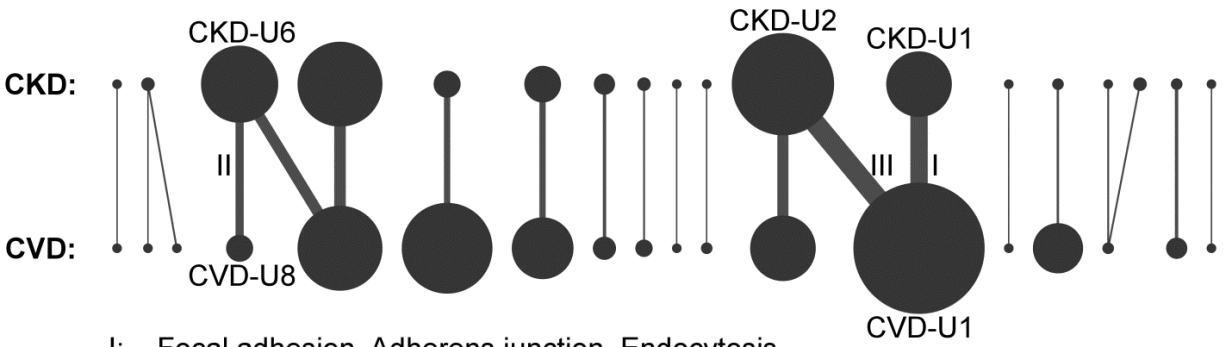


Figure 3: Molecular models for a) CKD and b) CVD. Each circle represents a segment (interpreted as molecular process) holding a subgraph (set of protein coding genes) as extracted from the disease term-specific subgraph. The circle diameter scales with the number of molecular features assigned. Edges between segments represent the aggregate relation of nodes across two segments. Segments being further discussed in the text are marked.



I: Focal adhesion, Adherens junction, Endocytosis
II: Retinol metabolism, Drug metabolism, Arachidonic acid metabolism, Linoleic acid metabolism, Metabolism of xenobiotics by cytochrome P450
III: Chemokine signaling pathway, Cytokine-cytokine receptor interaction

Figure 4: Interference of the CKD and CVD molecular model resulting in segments exhibiting significant overlap of molecular features, and edges scaling with the number of such shared features. Circle diameters scale with number of molecular features assigned to a segment. Roman number labels indicate specific enrichment of KEGG processes.