



universität
wien

DISSERTATION

Titel der Dissertation

Kidney Disease Interactome:

Systems Biology for analyzing Kidney Diseases

Verfasserin

Mag.rer.nat. Irmgard Mühlberger

angestrebter akademischer Grad

Doktorin der Naturwissenschaften (Dr.rer.nat.)

Wien, 2011

Studienkennzahl lt. Studienblatt: A 091 490

Dissertationsgebiet lt. Studienblatt: Molekulare Biologie

Betreuerin / Betreuer: Univ.Doiz.Dr. Bernd Mayer

*"Whereas the beautiful is limited, the sublime is limitless,
so that the mind in the presence of the sublime,
attempting to imagine what it cannot,
has pain in the failure but pleasure
in contemplating the immensity of the attempt."*

Immanuel Kant

Acknowledgements

I would like to take this opportunity to acknowledge and thank a number of people who got involved in this thesis in one way or another.

First and foremost, I wish to express my special gratitude to my family and friends for giving me assistance in any respect, particularly during the completion of this project.

I am furthermore heartily thankful to my supervisor, Bernd Mayer, whose encouragement, guidance and support from the initial to the final level of the working process, enabled this thesis.

Many thanks go to Rainer Oberbauer and Gert Mayer. I greatly benefited from their clinical expertise.

Moreover, it is my great pleasure to acknowledge the valuable support and help of Paul Perco, whose suggestions, comments and ideas guided me throughout this working process. His manners of providing orientation always helped me to get back on track during challenging times.

Also, my overall gratefulness is devoted to all of my colleagues. I want to seize this opportunity to express my true gratitude for the experience of being part of the emergentec team.

Finally, my deepest appreciation goes to Flo for his patience, love and endless support. He has been my anchor, throughout this work and beyond.

Index

1. Introduction.....	9
1.1 Systems Biology	9
1.1.1 The Evolution of High-throughput Technologies.....	9
1.1.2 Computational Systems Biology.....	11
1.1.3 Network Biology	12
1.1.4 Systems Biology in Disease	13
1.2 Data Sources	15
1.2.1 Omics Technologies	15
1.2.2 Literature Mining	21
1.3 Analysis Workflows	23
1.3.1 Sequential Workflows.....	23
1.3.2 Integrated Workflows.....	25
1.4 Applications.....	26
1.4.1 Acute Renal Failure/Transplantation.....	26
1.4.2 Chronic Kidney Disease.....	28
1.4.3 Cardiorenal Syndrome	29
2. Publications	31
2.1 Computational analysis workflows for Omics data interpretation. Bioinformatics for Omics Data: Methods and Protocols. 2011	31
2.1.1 The Thesis Author's Contribution.....	55
2.2 Biomarkers in renal transplantation ischemia reperfusion injury. Transplantation. 2009	57
2.2.1 The Thesis Author's Contribution.....	73
2.3 Impaired metabolism in donor kidney grafts after steroid pretreatment. Transpl Int. 2010.....	75
2.3.1 The Thesis Author's Contribution.....	103
2.4 Linking transcriptomic and proteomic data on the level of protein interaction networks. Electrophoresis. 2010	105
2.4.1 The Thesis Author's Contribution.....	129

2.5 Integrative bioinformatics analysis of proteins associated with the cardiorenal syndrome. Int J Nephrol. 2010	131
2.5.1 The Thesis Author's Contribution	154
2.6 Molecular pathways and crosstalk characterizing the cardiorenal syndrome. submitted to J Cell Mol Med.	155
2.6.1 The Thesis Author's Contribution	176
3. Discussion	177
3.1 Major Findings	177
3.1.1 Omics Workflows	177
3.1.2 Acute Renal Failure/Transplantation	177
3.1.3 Chronic Kidney Disease.....	179
3.1.4 Cardiorenal Syndrome	180
3.2 Outlook	182
4. Appendix.....	185
References	185
Abstract	191
Zusammenfassung	193
Curriculum Vitae	195

1. Introduction

1.1 Systems Biology

Systems Biology refers to a field in molecular biosciences that aims to understand molecular mechanisms of cells, tissues, or organisms by integrative analysis of multiple molecular and cellular components. However, since a system is not only the mere assembly of its components, a system-level understanding cannot be achieved by the study of singular molecules one by one and the focus of research has shifted from single elements to networks, from matters to states, and from structures to dynamics [1].

The following sections provide an overview of the different aspects and concepts of Systems Biology that build the basis for the concepts and methods used in the studies presented in this thesis.

1.1.1 The Evolution of High-throughput Technologies

The idea of understanding biological entities as dynamic systems is not new, but the availability of methods for investigating them as such led to a tremendous increase of research in this field as seen over the last decades. Along with the technical progress, the advance of high-throughput technologies opened up possibilities for a more global view on cellular processes. Large-scale data generation triggered the advent of a new domain which can be embraced by the term “omics”, and refers to the comprehensive analysis of a biological system on the respective level of observation, including genomics, transcriptomics, proteomics and many more (an overview of the different omics technologies is given in section 1.2.1). The enormous increase in data amounts is illustrated in figure 1 referencing the number of available sequences provided in the NCBI RefSeq database between 2004 and 2011 (available at <http://ftp.ncbi.nih.gov/refseq/release/release-statistics/>).

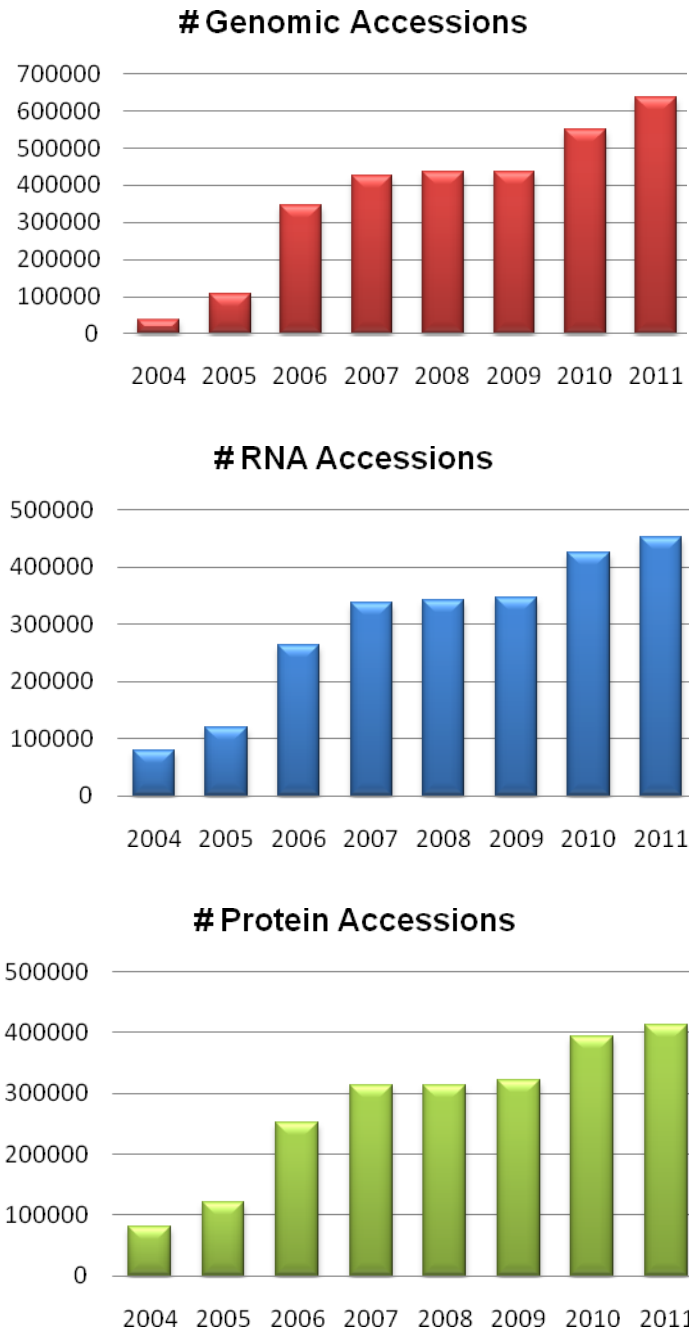


Figure 1: The histograms show the increase in number of accessions available in the NCBI RefSeq database between 2004 and 2011 for genomic, RNA, and protein sequences respectively.

The progress of high-throughput technologies has also implicated a shift from a traditional hypothesis-driven to a data-driven research. Data-driven or “top-down” approaches make use of an iterative cycle that starts with experimental data, followed

by data analysis and data integration, and ends with the formulation of hypotheses [2]. However, the 'traditional' way to define a hypothesis and design experiments for hypothesis testing cannot be completely replaced by data-driven approaches. Rüegg and colleagues [3] described a rather complementary relationship between omics data and hypothesis-driven research due to three major reasons: (i) profiles of omics-based studies have intrinsic limitations because of their descriptive nature, (ii) the integration of multi-level omics datasets has the potential to feed hypothesis-driven research, (iii) omics approaches may generate unexpected results that could not have been anticipated by hypothesis-driven research.

Certainly, the knowledge on the identity of the entirety of cellular components on a respective level of observation, as, for example achieved with the completion of sequencing of the human genome, has significantly contributed to advances in cellular biology. Nevertheless, the human genome has surprisingly few genes compared to far simpler organisms like *C. elegans*, opening the question where the difference in biological complexity comes from. Answers standing to reason are the interactions between genes, proteins and their regulatory mechanisms and can be addressed by the integration of data from different levels of observation. With the increase of data amounts and the rise in complexity of analysis strategies, the use of informatics techniques became necessary.

1.1.2 Computational Systems Biology

An important issue that arises from the generation of large quantities of data is their appropriate handling which concerns analysis, collection, classification, visualization, manipulation, storage, as well as dissemination of the acquired information. Thus, the integration of experimental and computational approaches became necessary and led to the emergence of the sub-discipline computational Systems Biology. To this point the need of interdisciplinary work became evident.

Basically, one can divide two groups of research in Systems Biology. One is the research on tools and algorithms for system-level studies. The other is research on system properties of specific biology, using the tools and algorithms developed. [1]. At a first glance, the former group is exclusively taken by computer scientists whereas the second one occupies the biologists. Nevertheless, scientists have to engage in interdisciplinary research collaborations to meet the demands of systems biology. At

minimum, computer scientists should acquaint themselves with the language of biology and biologists should understand the language of mathematics and computer modeling [4] in order to successfully take advantage of the possibilities that are provided by the integration of both disciplines.

Achievements of these combined efforts are reflected by the long list of bioinformatics tools that emerged during the last decades, ranging from statistical data processing to data annotation, data integration, and data management services. Remaining challenges concern the development of tools for automated workflow processing, allowing the use of multiple tools and the integration of data from different sources.

1.1.3 Network Biology

A major challenge in Systems Biology which inevitably demands a computational approach is the modeling of complex biological systems, for example the representation of relationships between cellular components as networks. The development of high-throughput techniques has allowed for the simultaneous interrogation of the status of cellular components [5], resulting in the emergence of comprehensive networks describing protein-protein interactions, metabolic reactions, signal transduction, and transcriptional regulation. Starting from the identification of small regulatory units (network motifs), networks can be built up to functional modules and in the end to large-scale organizational networks. The recognition of the modularity of many biological systems has brought remarkable insights into cellular organization. Since modules are defined as relatively small units with functional separation from other modules, they are manageable to undergo characterization. Higher-level properties of cells, such as their ability to integrate information from multiple sources, can be described by the pattern of connections among their functional modules [6].

Here again, in view of the temporally and spatially dynamic properties of biological systems, as well as of the obvious dependencies between the different types of networks, the need of integration of data from different levels of observation, including genes, proteins, or metabolites, becomes evident. Thus, the combination of the currently mostly separate layers of information in networks is demanded to enhance the understanding of cellular function [7]. In the course of our analyses towards

molecular characterization of kidney diseases, we made use of an integrated network approach that included parameters derived from multiple omics data and functional characteristics [8] (see section 2.1, 2.2, 2.3, 2.4, 2.5).

1.1.4 Systems Biology in Disease

Systems biology has also found its way into translational clinical research. Hallmarks of the emerging domain “systems medicine” are the establishment of new links between genes, biological functions, and a wide range of human diseases, thereby providing signatures of pathological biology and links to clinical research and drug discovery [9]. The use of high-throughput techniques is nowadays a common procedure for the identification of disease specific molecular signatures. Either directly linked to clinical outcomes or unsupervised processed and subsequently assessed for clinical trends, such signatures can serve as reference points for the identification of novel biomarker candidates. Since changes in gene or protein expression can often be detected before clinical symptoms arise, molecular markers have the potential for significantly improving risk assessment, diagnostic, and prognostic capabilities.

Unquestionably, genome-wide approaches had a significant impact on the development of analysis strategies and workflows for biomarker and drug target discovery. A prominent example is the “Human Disease Network” [10], a conceptual framework linking all genetic disorders (the human “disease phenome”) with the complete list of disease genes (the “disease genome”). This combined set of all known disease-gene associations provides a global view of the “diseasome” that significantly expands the traditional single-gene to single-disease approach (see Figure 2).

An insufficient understanding of the complex pathophysiology of many human diseases, including kidney dysfunction, is often the cause for a lack of early diagnosis strategies and efficient therapies. In response of this situation that points towards the need of a systems level understanding, the studies presented in this thesis aim to identify multiple aspects of diseases mechanisms by the integration of data from different sources that will be described in the following section.

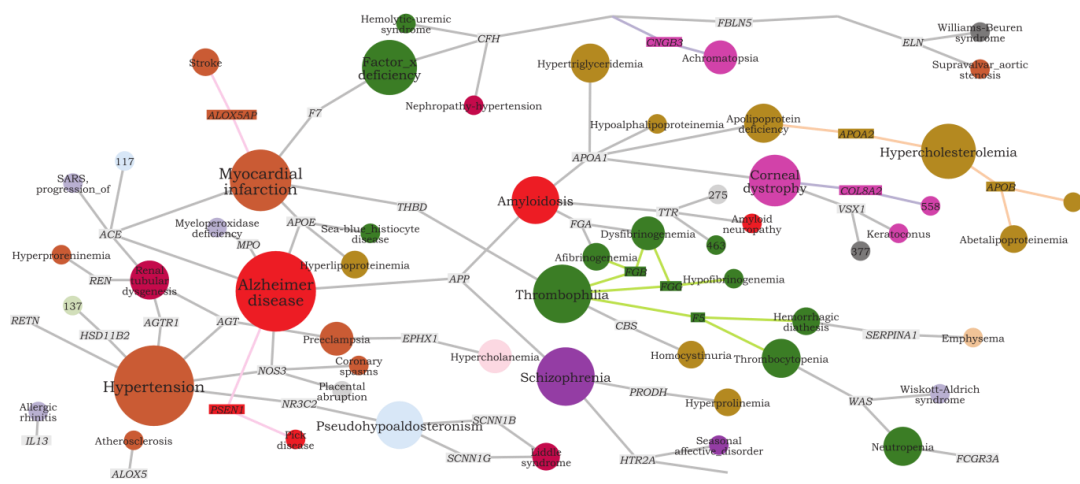


Figure 2: An excerpt of the “Human Disease Network” [10].

1.2 Data Sources

1.2.1 Omics Technologies

As outlined in section 1.1, the advent of high-throughput technologies has significantly contributed to advances in cellular biology. The generation of large amounts of unbiased data covering the totality of features on a respective level of observation allowed distinguishing more details of the cellular system. Each of the various “omes” listed in Table1 refers to one of these levels and is subject of study of the respective “omics” field.

Terms	Description	# Articles in Pubmed (April 2011)	First Year in Pubmed
Genome	The full complement of genetic information both coding and non coding in the organism	752488	1932
Transcriptome	The population of mRNA transcripts in the cell, weighted by their expression levels	62903	1997
Proteome	The protein-coding regions of the genome	18469	1995
Metabolome	The quantitative complement of all the small molecules present in a cell in a specific physiological state	1549	1998
Interactome	List of interactions between all macromolecules in a cell	750	1999
Secretome	The population of gene products that are secreted from the cell	464	2000
Glycome	The population of carbohydrate molecules in the cell	163	1999
Phenome	Qualitative identification of the form and function derived from genes, but lacking a quantitative, integrative definition	152	1995
Physiome	Quantitative description of the physiological dynamics or functions of the whole organism	108	1997
Orfeome	The sum total of open reading frames in the genome, without regard to whether or not they code; a subset of this is the proteome	67	2002
Cellome	The entire complement of molecules and their interactions within a cell	36	2002
Fluxome	The population of proteins weighted by their fluxes	34	1999
Regulome	Genome-wide regulatory network of the cell	20	2004
Translatome	The population of mRNA transcripts in the cell, weighted by their expression levels	9	2001
Transportome	The population of the gene products that are transported; this includes the secretome	9	2004
Localizome	The localization of various proteins, both in terms of cell type and subcellular compartments	6	2001

Ribonome	The population of RNA-coding regions of the genome	4	2002
Morphome	The quantitative description of anatomical structure, biochemical and chemical composition of an intact organism, including its genome, proteome, cell, tissue and organ structures	3	1996
Operome	The characterization of proteins with unknown biological function	1	2002
Functome	The population of gene products classified by their functions	1	2001
Foldome	The population of gene products classified by their tertiary structure	1	2009
Pseudome	The population of pseudogenes in the cell	0	-
Unknome	Genes of unknown function	0	-

Table 1: List of the different “omes”. Given are the descriptions, the number of articles found in Pubmed and the year of its first appearance. (The table is an updated version of the “Omes Table” available at <http://bioinfo.mbb.yale.edu/what-is-it/omes/omes.html>.)

Along with the explosion of omics data amounts, a multitude of public databases providing data of different omics tracks came up and the need for standards for data annotation and exchange arose. The next sections provide an overview on the most common omics technologies, together with examples of available databases and standards used.

Genomics

Genomics is classically divided into two areas, namely structural and functional genomics. Whereas the target of research in the former is DNA, functional genomics, or the “post-genomic area”, deals with functional aspects of DNA and also includes transcripts, proteins and metabolites which will be discussed later. Structural genomics includes DNA sequencing, as well as studies on DNA complexity, DNA variability, DNA genotyping, DNA organization within the cell, and DNA modification [11].

Sequencing methods were the first high-throughput techniques developed and the first genome, a single-stranded bacteriophage, was completely sequenced in 1977 [12]. Today, the NCBI Genome Project database (<http://www.ncbi.nlm.nih.gov/genomes>) holds 1014 completed genome sequencing projects, and further 938 projects are in progress (status April 2011).

The currently available sequencing methods were recently reviewed by Kircher et al. and include Sanger capillary sequencing, pyrosequencing, reversible terminator chemistry, sequencing-by-ligation, and virtual terminator chemistry [13].

The effective use of large scale data requires the establishment of standardized methods that support exchange, annotation, archiving, and mining of existing data sets. In the last years, considerable efforts were made by the scientific community concerning this matter and resulted in a number of standards with different scopes, ranging from reporting, data exchange, terminology to physical and data analysis standards, developed by several institutions. In case of genomics, the most common reporting standards include MIGS/MIMS (Minimum Information about a Genome/Metagenomic Sequence/Sample, developed by the Genomic Standards Consortium), or MINSEQE (Minimum Information about a high-throughput Nucleotide Sequencing Experiment, developed by the Microarray Gene Expression Data Society).

A list of common publicly available genomic sequence databases is provided in table 2.

Name	Web Link
NCBI Genome database	http://www.ncbi.nlm.nih.gov/genome
NCBI Reference Sequence database	http://www.ncbi.nlm.nih.gov/RefSeq/
EMBL Nucleotide Sequence database	http://www.ebi.ac.uk/embl/
Ensembl Genomes	http://www.ensemblgenomes.org/

Table 2: Common publicly available genome sequence databases.

Transcriptomics

Transcriptomics usually refers to the large scale analysis of gene expression patterns. The first lines of transcriptomic studies can be dated back to 1965 where the sequence of the first RNA molecule was determined [14]. Further milestones were the introduction of Northern blots, Real-time PCR, and differential display with relatively low experimental throughput. In the nineties, the development of SAGE (Serial Analysis of Gene Expression) and microarrays has sounded the bell for the era of genome-wide transcriptomics, in practice covering the protein coding genome. Over the years, gene expression profiling techniques have continuously advanced. Tiling and exon arrays

are available and the advent of next-generation sequencing offered the possibility of large scale transcriptomics at a single nucleotide resolution [15].

However, the most prominent transcriptomics technologies are still DNA microarrays. The basic principle of microarrays is base-pairing which is experimentally achieved by the hybridization of targets to gene specific sequences that are immobilized on a solid state matrix. Basically, all arrays employ the same four components: (i) target labeling, (ii) target-probe hybridization, (iii) detection and (iv) data analysis [16]. The type of targets to use is determined by the immobilized probes which are mostly cDNA sequences or oligonucleotides. Furthermore, DNA arrays can be classified into one-channel and two-channel arrays, reflecting the difference of hybridization of both samples to be compared on one array or on two arrays. Former provides absolute values on mRNA concentration whereas signals of two-channel arrays represent relative measurement of gene expression.

A description of microarray data processing and analysis is partly given in section 1.3 and in great detail in section 2.1.

The most common reporting standard for microarray experiments is MIAME (Minimum Information About a Microarray Experiment, developed by the Microarray Gene Expression Data Society). It aims to enable the interpretation of the results of an experiment unambiguously and potentially to reproduce the experiment. The six most critical elements contributing towards MIAME are: (i) raw data, (ii) processed data, (iii) sample annotation including experimental factors, (iv) experimental design, (v) array annotation and (vi) laboratory and data processing protocols [17].

Most of the common databases for microarray data are compliant with the MIAME standards. Table 3 provides a list of public repositories.

Name	Web Link
NCBI Gene Expression Omnibus	http://www.ncbi.nlm.nih.gov/geo/
EMBL Array Express	http://www.ebi.ac.uk/arrayexpress/
Oncomine (cancer transcriptome profiles)	https://www.oncomine.org/
Nephromine (kidney transcriptome profiles)	http://www.nephromine.org/

Table 3: Common microarray data repositories.

Proteomics

Proteomics comprises the systematic functional and structural analysis of proteins. Since the completion of the human genome project, scientists aim to annotate the genome with protein-level information [18,19]. Considering the large number of factors that determine individual protein concentrations, the challenges of these projects become evident. These include the controls on the transcription of genes, the codon usage, the rates and extent of post-translational modification, nature and abundance of proteins with which the gene product interacts, substrate levels and rates of proteolytic degradation [20]. Actually because of the inadequate prediction of protein abundance from mRNA concentrations, the direct measurement of protein expression is demanded.

Identification and quantification of proteins is usually a two-step procedure, starting with the separation of the isolated protein mixtures, followed by the quantification and identification of the individual components using mass spectrometry or similar approaches. The classical method which is still most widely used for protein separation is two-dimensional gel electrophoreses (2DE). In 2DE, proteins are first separated by isoelectric focusing and then further resolved by mass using SDS–PAGE. Additional strategies commonly in use are chromatographic purification methods that separate proteins based on their physiochemical properties. Methods for quantification of proteins include comparative 2DE approaches, in vivo metabolic labeling, or isotope-coded affinity tagging (ICAT) [21]. Mass spectrometry can provide quantitative, as well as qualitative information. By definition, a mass spectrometer consists of an ion source, a mass analyser that measures the mass-to-charge ratio (m/z) of the ionized analytes, and a detector that registers the number of ions at each m/z value [22]. Coupled MS including protein fragmentation determines the molecular m/z ratio of peptides, which are then used to identify the predicated proteins using web-based search engines such as MASCOT and PROFOUND [23].

As biological functionality is largely driven by the interaction of biologically active molecules, the identification of protein—protein interactions poses a further important field in proteomics research. Commonly used methods for protein interaction determination include the yeast two hybrid systems, protein arrays and affinity chromatography.

In analogy to the MIAME standard for microarrays, the Human Proteome Organization Proteomics Standards Initiative (HUPO-PSI) has introduced the MIAPE standard for

reporting proteomic experiments. HUPO-PSI also defines, among others, a standard for the documentation of protein interactions called MIF (Molecular Interaction Format).

Table 4 provides an excerpt of the list of the variety of publicly available data repositories and resources holding protein related information as sequence, structure, or interaction data. Comprehensive databases on tissue or disease specific proteins are comparably rare which may be due to modest numbers of samples and the difficulty of merging data from more than one study across different analytical platforms. One example is the Human Urinary Proteome database [24] that was in this thesis used for the extraction of chronic kidney disease specific proteins for the cross-omics study presented in section 2.4. This database holds the information about protein abundance of 3687 human urine samples (status as of September 2009) that were collected from patients covering a wide spectrum of different pathophysiological conditions, among them renal disorders, as well as from healthy controls.

Name	Web Link
Sequence	
NCBI Protein database	http://www.ncbi.nlm.nih.gov/protein/
NCBI Reference Sequence database	http://www.ncbi.nlm.nih.gov/RefSeq/
UniProtKB Protein knowledgebase	http://www.uniprot.org/
Structure	
RSCB Protein Databank	http://www.pdb.org/
ExpASY Database of annotated 3D Images	http://expasy.org/sw3d/
Protein Interactions	
EMBL Protein Interaction Database	http://www.ebi.ac.uk/intact/
Online Predicted Human Interaction Database	http://ophid.utoronto.ca/
Biomolecular Interaction Network Database	http://bind.ca/

Table 4: Common publicly available proteomic databases

Metabolomics

The Human Metabolome Project started in 2004 and, although the Human Metabolite Database holds nearly 8600 compounds, the identification of the human metabolome is still far from complete [25,26]. One of the great challenges towards completeness is the analytical bias due to chemical properties of different compound classes. Metabolic

profiling by contrast focuses at the quantitative analysis of a set of pre-defined metabolites belonging to a class of compounds or members of particular pathways. A further subsection of metabolomic analyses is target oriented and aims at a quantitative analysis of substrate or product metabolites of a single target protein [27].

The basic principles of metabolite identification and quantification are similar to those of proteomics. Separation methods are mostly chromatographic or electrophoretic techniques. Mass spectrometry or Nuclear Magnetic Resonance (NMR) spectroscopy are usually the methods of choice for the detection of the metabolites.

The CIMR (Core Information for Metabolomics Reporting) standard specifies the minimal guidelines reporting metabolomics work and was introduced by the Metabolomics Standards Initiative (MSI).

Beside of the Human Metabolome Database which is currently the most complete and comprehensive curated collection of human metabolite data, there exist a number of resources containing information on small molecules. Examples are given in table 5.

Name	Web Link
Human Metabolome Database	http://www.hmdb.ca/
KEGG Ligand database	http://www.genome.jp/kegg/ligand.html
NCBI PubChem	http://pubchem.ncbi.nlm.nih.gov/

Table 5: Examples for public repositories of small molecules.

1.2.2 Literature Mining

Text-mining in molecular biology is defined as the automatic extraction of information about genes, proteins and their functional relationships from text documents [28].

The increasing number of electronically accessible publications has opened the door for efficiently taking advantage of the results from the combined efforts of the scientific community that are provided within the literature.

Basic resources for biomedical literature mining tools are databases like PubMed which currently holds about 20 million abstracts. The development of textual databases and

ontologies that catalog and organize terms to assist authors in consistent use of domain specific terminology has significantly improved text mining approaches. Furthermore, databases providing training text collections for machine learning approaches have been constructed [29].

Of special interest in the context of gene-disease associations is the co-appearance of disease concepts and gene names within one and the same publication which gives information about relevant genes for a certain disease phenotype. Publications indexed in PubMed are annotated with Medical Subject Headings (MeSH) maintained by the U.S. Library of Medicine which are organized in a hierarchical structure of sub- and super-categories. Thus, the MeSH terms in the disease category can be used for a paper-disease mapping. Unfortunately, this framework has one considerable drawback. Diseases that are not part of the MeSH universe, for example the cardiorenal syndrome, cannot be handled and a paper-disease mapping must be obtained by free-text search. A subsequent paper-gene mapping can be obtained, for example, from a NCBI curated file (<ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2pubmed.gz>), that represents the logical equivalent of what is reported as Gene/PubMed links available in Gene's and PubMed's links menus on the NCBI homepage. Further tools for the automatic detection of protein and gene mentions from the literature include the GAPSCORE [30] or ABGENE [31] systems.

1.3 Analysis Workflows

The basic goal of analysis workflows applied to any forms of omics data is to transform raw data sets into interpretable information and knowledge on a biological level [32]. In the studies presented in this thesis, we made use of sequential analysis procedures, as well as of integrated approaches with focus on protein interaction networks. Section 2.1 provides a detailed description of both forms of workflows together with a list of available resources and tools, supplemented with an example workflow on a gene expression dataset. An overview on the principle concepts of transcriptomics data analysis is given in the following.

1.3.1 Sequential Workflows

A sequential analysis workflow follows a step-by-step procedure starting from the raw datasets and ending in a functional interpretation where identified features are embedded in their biological context, allowing the generation of hypotheses. The main steps are usually (i) raw data processing, (ii) statistical analysis, and (iii) functional analysis.

Data Preprocessing

The need of microarray data preprocessing arises from the fact that intensity values not only reflect actual mRNA concentrations but are influenced by several non-biological factors. Examples are variations in the array manufacturing process, the preparation of the biological sample, the hybridization of the sample to the array, or the quantification of the spot intensities [33]. In order to ensure the comparability of arrays within an experiment, removal of the estimated background signal and normalization between arrays are usually performed. The application of filter routines, e.g. based on the number of missing values or marginally expressed genes, leads to a reasonable reduction in complexity of microarray data in terms of gaining usable information at least from a statistical perspective.

Statistical Analysis

Most of the microarray experiments aim at the detection of quantitative differences in gene expression between two groups of samples representing two conditions (e.g. case/control). A first impression of array grouping can be obtained by the visualization of array clustering based on their proximity (defined by a distance function) to each other following unsupervised clustering approaches. Resulting dendrograms that reflect the initially considered grouping give a lead to a succeeded experiment whereas controversial outcomes may indicate a systemic bias that can be due to experimental issues as, for example, different sample preparation, uneven hybridization, or different array batches.

The next step is the identification of differentially expressed genes between the sample groups. A simple method is the calculation of fold-changes, but it has been shown that the fold-change criterion alone is unreliable because statistical variability is not taken into account [34,35]. More sophisticated procedures involve test statistics that assign a statistical significance score (p-value) to each gene. Considering the large number of comparisons that are made for each probe on the array, a correction for multiple testing is indispensable for the reduction of false positive findings. Furthermore, a p-value cut-off above which biologically meaningful information is expected has to be defined. Depending on the chosen cut-off, a more or less manageable list of differentially expressed list is available for interpretation in the given biological context.

Functional Analysis

A decisive step that ensures the proper mapping of genes to functional categories is their consistent annotation to unique identifiers. The list of already established biological identifiers is long and different functional annotation tools often require different identifiers.

The identification of statistically enriched or depleted functional categories follows the principal foundation that if a biological process is perturbed in a given study, the functionally linked genes (on the level of proteins as effector molecules) should have a higher potential to be selected as a relevant group by the high-throughput screening technologies [36]. This assumption can be expanded to different levels of functional relationships including molecular functions or pathways.

Further approaches for functional analysis are, among others, protein interaction networks, the detection of co-regulation, tissue specific expression, or protein subcellular location.

1.3.2 Integrated Workflows

Following the rationale that the cell is an integrated system and its biological mechanisms cannot be fully described by the observation of single layers, the approach of integrating multiple omics data and different functional characteristics into analysis procedures reflects the spirit of Systems Biology and became increasingly popular along with the advent of suitable technologies.

Major challenges in the field of integrative bioinformatics address the usability of heterogeneous data since most data sources still exist in isolation where each source has its own specialization and focus. In many cases, databases lack links to each other, even when they are providing data about the same entities [37,38].

Many of the integrated approaches are based on interaction networks that represent functional dependencies derived from the input of multi-level data. In addition to physical interactions between biological entities, such networks include indirect associations such as co-regulation or shared pathway memberships that are equally important for a complete understanding of biological systems [39-41].

1.4 Applications

In the present thesis, the concepts discussed in the previous sections were applied on various forms of kidney disease, spanning from acute renal failure in the transplant situation, further to chronic kidney disease, and finally to cross-organ analysis, namely the cardiorenal syndrome.

A short summary of the basic facts on kidney function and structure are given in the following:

Basically, the kidney performs two main functions: (i) the organ participates in the maintenance of a constant extracellular environment by the excretion of metabolic waste products, electrolytes and water, and (ii) the organ secretes hormones involved in hemodynamic regulation, production of erythrocytes, and mineral bone metabolism [42].

The functional unit of the kidney is the nephron which consists of the renal corpus (glomerulus and Bowman's capsule) which is responsible for filtering and the renal tubule (proximal tubule, loop of Henle, distal tubule) functioning as absorption and secretion apparatus.

The following sections provide an overview on the pathophysiology of different types of kidney disease, namely acute renal failure, chronic kidney disease, and the cardiorenal syndrome.

1.4.1 Acute Renal Failure/Transplantation

Acute renal failure (ARF) is characterized by the abrupt decline in glomerular filtration rate [43] as a result of vasoconstriction, hypoxia, ischemia, or the usage of nephrotoxic substances. It affects 25% - 30% of patients in the intensive care unit and 3% - 7% of patients admitted to the hospital [44].

Until a few years ago, a consensus definition of ARF was lacking. In 2004, the ADQI (Acute Dialysis Quality Initiative) group proposed the RIFLE criteria for staging ARF patients, the initials reflecting the terms Risk, Injury, Failure, Loss and End Stage in relation to kidney function [45].

The traditional etiological classification divides in prerenal, intrarenal, and postrenal causes. Prerenal ARF is the most common type and can be caused by volume loss,

decreased cardiac output, neurogenic dysfunction, or vessel diseases. Intrarenal ARF is intrinsic and a response to tubular, glomerular, interstitial, or vascular injury. Postrenal ARF refers to the consequences of the obstruction of outflow tracts of the kidney [46].

ARF frequently appears in the post-transplant situation in context of a delayed graft function. Risk factors include donor age and cause of death or the duration of cold ischemia with consequences leading to a reduced long-term allograft survival. Since intrinsic donor factors are among the main contributors to post-transplant ARF, including the autonomous cytokine storm after brain death and hemodynamic instability, the use of cadaveric donor organs has significant impacts on graft function. Several studies report a highly increased risk for post-transplant ARF in this patient group [47,48]. Changes in gene expression that distinguish living and cadaveric donor organs could be found in the functional categories inflammation, complement and coagulation, apoptosis, and cell adhesion [49]. The results of a double-blinded, randomized, controlled trial of steroid or placebo infusion into deceased donors and the consequences on graft function are presented in section 2.3.

Tubular and vascular damage in the donor organ after cold ischemia but before transplantation is associated with subsequent ischemic reperfusion injury (IRI) and an additional contributor to delayed graft function. Biomarkers for the detection of early injury, determination of graft quality, and prediction of graft outcome are demanded.

A routinely used marker for the diagnosis of ARF is the concentration of creatinine in blood which rises with the progression of glomerular filtration deficiency. This has been the method of choice for ARF diagnosis for nearly 60 years, but its limitations regarding the delayed rise in serum creatinine with respect to the decrease of the glomerular filtration rate, and the lack of specificity and sensitivity are evident [50,51]. Alternative biomarker candidates include Cystatin C, Neutrophil Gelatinase-Associated Lipocalin (NGAL), Interleukin-18 (IL18), and the Kidney Injury Molecule 1 (KIM1) [52,53], but further validation and trials are required to substantiate the utility of these markers. A review on biomarkers in renal transplantation IRI is provided in section 2.2 of this thesis.

1.4.2 Chronic Kidney Disease

The prevalence of chronic kidney diseases (CKD) in the general population is dramatically high with 11% of adults suffering from a reduced kidney function [54]. Loosely speaking, CKD is the progressive loss of kidney function over a period of months or years. According to the guidelines of the Kidney Disease Outcome Quality Initiative (KDOQI), CKD can be divided into 5 stages with respect to the decline in glomerular filtration rate, ranging from normal to relatively high GFR (stage 1) to kidney failure and the need of renal replacement therapy (stage 5) [54].

The most common causes for reaching a chronic state of kidney diseases include diabetes mellitus, hypertension, glomerulonephritis, interstitial nephritis, and low flow states (hypoperfusion) [55]. However, independent of the origin, most of the renal diseases that are the starting point for CKD begin with glomerular dysfunction. With ongoing disease progression, the glomerular injury expands to the tubulointerstitium, the connective tissue surrounding the renal tubule, leading to nephron loss and fibrotic lesions. The loss of functioning nephrons in turn causes an increased workload for the remaining nephrons with glomerular hypotension as consequence, thereby generating an ongoing vicious circle of progressive kidney damage [56].

An independent marker of worsening of kidney function is the loss of proteins in the urine that can be either due to a reduced glomerular filtration or a low absorption of the proximal tubulus. Unfortunately, the increased passage of proteins across the glomerular capillary barrier is not solely a consequence of renal injury but contributes to further disease progression. The exposure of tubular cells to plasma proteins further induces damage by the stimulation tubular chemokine expression and complement activation, leading to inflammatory cell infiltration in the interstitium and subsequent fibrogenesis [57]. Moreover, an increased excretion of albumin may result in hypoalbuminaemia which can be linked to an impairment of immune function.

In view of the fact that the estimation of the glomerular filtration rate by measuring the creatinine clearance is limited in its diagnostic and prognostic value (see 1.4.1), several efforts have been made to identify more accurate markers. Examples include Neutrophil Gelatinase-Associated Lipocalin (NGAL), Kidney Injury Molecule 1 (KIM1), Urinary liver-type fatty acid binding protein (urinary L-FABP), connective tissue growth factor (CTGF), transforming growth factor β (TGF β), or urinary mRNAs [52,58]. Next to the identification of single molecular biomarkers, approaches tending to find whole panels of markers, not least owing to advantages of high-throughput technologies,

promise more specificity for future diagnosis and prognosis of CKD. Section 2.4 presents an integrated approach for characterizing CKD mechanisms by the joint interpretation of transcriptomics and proteomics datasets.

It is known that mortality in patients with CKD is mainly due to adverse outcomes rather than to the kidney failure per se, with the leading causes of death being cardiovascular diseases. The next section discusses the pathophysiological connection between the kidney and the cardiovascular system, referring to the cardiorenal syndrome.

1.4.3 Cardiorenal Syndrome

Chronic kidney disease is encountered by a significant increase of cardiovascular complications. In dialysis patients the prevalence of cardiovascular disease (CVD) and the mortality due to CVD is around 10 to 30 times higher than in the general population [59]. The pathophysiological state of combined kidney and cardiovascular dysfunction is termed the cardiorenal syndrome.

Basically, the CRS can be classified into 5 subtypes, depending on the origin of damage (either the cardiovascular system or the kidney) and the course of disease (either acute or chronic) [60,61]. Figure 3 provides an overview on interactions referring to CRS types 2 and 4 (chronic cardiorenal syndrome and chronic renocardiac syndrome).

As can be seen, the significant impact of consequences of renal impairment on cardiovascular function, including the development of anemia, a fluid overload and the systemic presence of uremic toxins, become already evident in early stages of CKD. However, the main risk factors for cardiovascular events, like hypertension, dyslipidemia, or chronic inflammation, appear in the course of progressed CKD and are significantly increased in the cohort of patients on dialysis treatment [63-65]. In turn, low cardiac output, possibly coupled to genetic or acquired risk factors, has negative effects on kidney function and, if reaching a chronic state, leads to sclerosis and fibrosis.

Hormone mediated hemodynamic dysregulation also plays a decisive role in CRS formation. The renin-angiotensin and natriuretic peptide system have counterbalancing effects on renal and cardiovascular function through their opposing actions on vascular tone and sodium and water balance as well as cellular hypertrophy and fibrosis.

Angiotensin-converting enzyme or vasopeptidase inhibitors were shown to provide important end-organ protection in CRS [66].

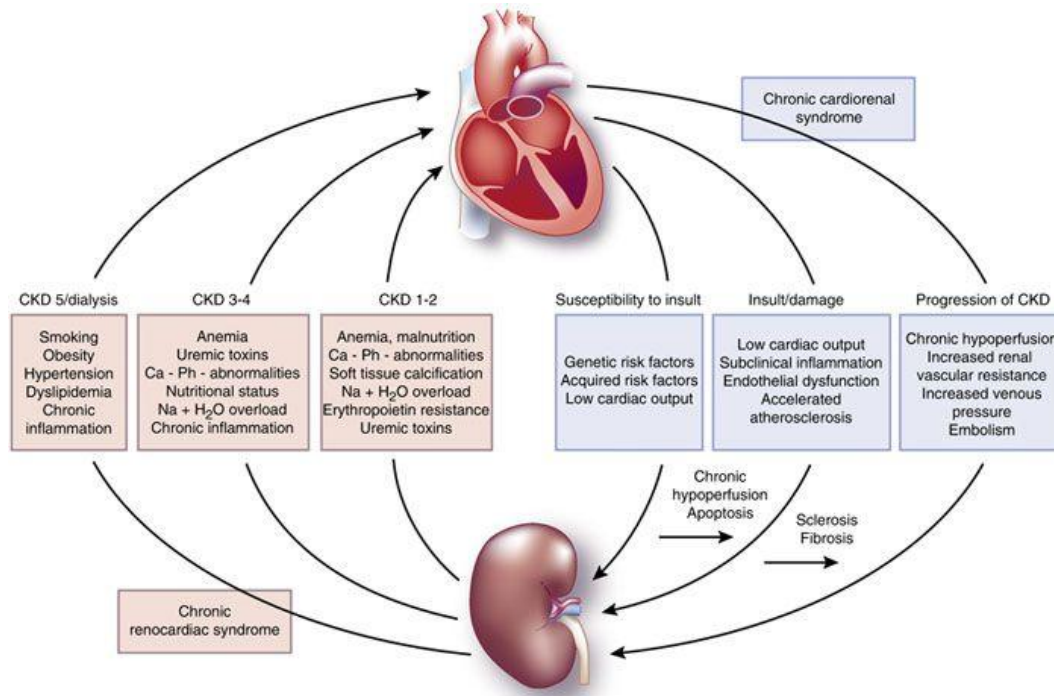


Figure 3: Overview on cardio-renal interactors and risk factors, taken from [62].

The complex characteristics of the CRS impose a new challenge in identifying and treating patients with CVD in early stages of CKD towards improving outcome. So far there is no clear understanding of the molecular pathways interlinking kidney failure and cardiovascular complications, concomitantly impeding the identification of biomarkers for identifying the risk of CVD in CKD patients. Sections 2.5 and 2.6 present two integrative studies analyzing proteins, pathways and the molecular crosstalk on the interface between the kidney and the cardiovascular system.

2. Publications

**2.1 Computational analysis workflows for Omics data interpretation.
Bioinformatics for Omics Data: Methods and Protocols. 2011**

Computational Analysis Workflows for Omics Data Interpretation

Irmgard Mühlberger, Julia Wilflingseder, Andreas Bernthaler, Raul Fechete,
Arno Lukas, and Paul Perco

Published in: Methods Mol Biol. 2011;719:379-97.

ABSTRACT

Progress in experimental procedures has led to rapid availability of Omics profiles. Various open-access as well as commercial tools have been developed for storage, analysis, and interpretation of transcriptomics, proteomics as well as metabolomics data. Generally, major analysis steps include data storage, retrieval, preprocessing and normalization, followed by identification of differentially expressed features, functional annotation on the level of biological processes and molecular pathways, as well as interpretation of gene lists in the context of protein-protein interaction networks. In this chapter we discuss a sequential transcriptomics data analysis workflow utilizing open-source tools, specifically exemplified on a gene expression dataset on familial hypercholesterolemia.

Key Words: Omics data analysis; bioinformatics workflow; transcription factor; protein network; data interpretation.

1. INTRODUCTION

High-throughput methods in molecular biology research, and in particular microarray technologies and mass spectrometry have led to the quantitative assessment of thousands of features on the level of the genome, transcriptome, proteome, and metabolome, resulting in the accumulation of a massive amount of data. Microarray technologies, initially restricted to applications in research, have in the meantime found its way into the clinic, e.g. referring to the MammaPrint microarray-based test system cleared by the FDA in early 2007 for the prognosis of breast cancer patients [1]. Next to basic research and molecular diagnostics, Omics procedures are also used for toxicological profiling as well as for drug discovery research in the hunt for novel therapeutic targets, just to give examples.

With these well established methodologies and standardized protocols for experimental processing in hand the emphasis of research in recent years has been on the analysis of high-throughput data and results interpretation [2]. Analyses steps include data storage, data annotation, data preprocessing and normalization, followed by explorative and statistical analyses, functional interpretation, and hypothesis generation. For all these different steps open-source tools are available and databases storing Omics raw data have been vigorously populated.

In this chapter we address computational analysis workflows for the interpretation of Omics data. We provide links to databases, tools and websites, discuss their applicability, and navigate through the analysis process on a given example dataset on gene expression profiles of monocytes from patients with familial hypercholesterolemia.

2. MATERIALS

2.1. Omics Data Repositories

Public databases provide genomics and proteomics data for a wide range of cells, tissues and diseases (**Table 1A**). Open-access repositories for microarray data are e.g. the ArrayExpressDB hosted by European Bioinformatics Institute (EBI) [3], the Gene Expression Omnibus developed at the National Center for Biotechnology Information (NCBI) [4], or the Stanford Microarray Database (SMD) [5]. One of the most comprehensive collections of proteomics data is provided by SWISS 2-D PAGE hosted by the Swiss Institute of Bioinformatics [6,7] (see **Note 1**).

Standards for data annotation and exchange of microarray data have been introduced by the Microarray Gene Expression Data (MGED) Society. The Minimum Information About a Microarray Experiment (MIAME) guidelines describe the minimum information needed for revising and interpreting results of a microarray-based experiment palpably [8].

2.2. Data Preprocessing

A sequence of data preprocessing steps is required for the analysis of abundance data e.g. from gene expression or protein profiling (**Table 1B**). Background correction and normalization of the data are the first steps to clear the impact of non-biological influences potentially arising from different array batches used or from varying intensities of different dyes. Frequently used background correction methods are the Robust Multi-array Average (RMA) method [9] or MAS 5.0 from the Affymetrix Microarray Suite [10]. Normalization techniques are Quantile Normalization (RMA), Invariant Difference Selection (IDS) [11], and dChip [12]. Further preprocessing is particularly important for gene expression data to achieve a reduction of data complexity. Filter routines focus on the elimination of entries which are probably invalid and will not contribute to informative results. One possible filter is to remove all objects for which the number of missing values over all experiments (arrays) performed exceeds a certain threshold. Missing values may be a problem caused by improper resolution, image corruption, or physical defects. Methods for handling missing values span from simple row average estimates to more sophisticated approaches e.g. based

on K-nearest-neighbor replacement [13], Bayesian variable selection [14], least squares replacement [15], or a combination of above mentioned procedures [16].

Preprocessing of proteomic MS data aims to identify a list of m/z peak values to be directly used for further analyses. Analyses steps include background correction, filtering, noise estimation, peak detection and spectral alignment algorithms [17-22]. Nie et al summarized current applications of statistics in several stages of global gel-free proteomic analysis by mass spectrometry [23]. For protein identification based on m/z data several resources are available as e.g. MASCOT [17].

After normalization issues are resolved the annotation of Omics features is essential. The SOURCE tool from the Stanford Genomics Facility [18] or the GeneCards system from the Weizmann Institute of Science [19] are commonly used annotation databases/tools for DNA/mRNA and protein sequences.

2.3. Identification of Differentially Expressed Genes and Proteins

For the evaluation of differentially expressed genes/proteins several methods based on test statistics are in use (**Table 1C**). A straightforward method is the Student's t-test determining the significance of differences between distributions of expression levels combined with computation of the fold change. The correction for multiple testing is pivotal for the analysis of Omics data in order to reduce the number of false positive findings. A very stringent correction method is the Bonferroni correction, whereas less conservative methods are based on permutations e.g. realized by the maxT and minP method as described by Westfall and Young [20]. Such permutation and resampling methods are described in detail by Dudoit et al. [21] and Ge et al. [22]. Implementations of these algorithms can be found in the *multtest* Bioconductor package of the R statistics environment [24,25]. Bootstrap and Jackknife procedures, both using randomly drawn subsets of the whole dataset, further strengthen the statistical findings and lower the susceptibility to outliers [26]. Significance Analysis of Microarrays (SAM) is also based on data permutation but controls the false discovery rate (FDR), defined as the percentage of genes identified as significant with respect to the number of features identified as relevant by chance [27]. This method is widely accepted in microarray analysis. SAM is available as stand-alone package and is also implemented in the MultiExperiment Viewer (MeV) developed at The Institute for Genomic Research (TIGR) [28].

2.4. Functional Annotation and Pathway Enrichment Analysis

One approach for functional grouping of genes or proteins identified as relevant from a statistical viewpoint is realized by utilizing gene ontologies (GO), categorizing proteins according to their molecular functions, cellular components, and biological processes (**Table 1D**). Another classification system is the PANTHER (Protein ANalysis THrough Evolutionary Relationships) ontology [29]. Generally, ontologies are controlled vocabularies and can be represented as acyclic, directed graphs where each ontology category can have one or more parent and sub terms. Statistical tools exist to identify enriched or depleted categories for a list of genes or proteins of interest [30]. One of these tools is DAVID (Database for Annotation, Visualization and Integrated Discovery) [31].

Pathway databases like the one from the Kyoto Encyclopedia of Genes and Genome (KEGG) [32] complement the functional ontologies and can give even more information on the interplay of gene and proteins. Other pathway databases describing metabolic networks and signaling transduction cascades are the BioCarta, the PANTHER pathway database [29], or Reactome [33]. KEGG spider provides a robust analytical framework for interpretation of gene lists in the context of a global gene metabolic network [34] (**Table 1E**).

2.5. In-silico Promoter Analysis

Transcription factors are key elements in the regulation of transcription exerting their function by binding to the promoter region of a gene as well as to regulatory elements further away from the transcription start site (**Table 1F**). JASPAR is a database holding binding site matrices for specific transcription factors which can be used by pattern matching algorithms in order to scan genomic sequences for potential transcription factor binding sites (TFBS) [35]. The JASPAR Core database provides a curated, non-redundant set of binding profiles from experimentally defined transcription factor binding sites for eukaryotes reported in the literature.

For a given list of differentially regulated genes or proteins the search for enriched TFBS in the regulatory regions becomes feasible. oPOSSUM is a database that contains pre-calculated transcription factor binding sites in the regulatory regions of human genes that can be used in order to identify enriched transcription factors in a set of deregulated genes [36]. The regulatory regions of human genes are identified searching for conserved regions in the mouse genome (phylogenetic footprinting) using

different stringency criteria. The oPOSSUM tool uses transcription factor binding sites as stored in the JASPAR database.

2.6. Integrated Approaches

Besides sequential workflows following a step-by-step analysis several integrated approaches exist (**Table 1G**). One example is STRING, provided by the European Bioinformatics Institute (EBI) which aims to present genes directly or indirectly related to a query gene [37,38]. The basis of STRING is a protein network obtained from integrating high-confidence data, high-throughput experiments, and computationally derived data for more than 2.5 million proteins occurring in 630 organisms. Information is integrated over organisms and the respective proteins are represented as clusters of orthologous groups. STRING currently integrates protein interactions, co-expression data, literature co-occurrences, genomic context encoded by conserved genomic neighborhoods, gene fusion events, and phylogenetic co-occurrences. For each pair of proteins STRING pre-computes a detailed measure of evidence based on each available data source for describing the association between the two proteins. These sub-scores are combined to represent an evidence score. A STRING query is performed by entering a gene name, protein name or a protein sequence, or a list of identifiers or sequences. As a result STRING shows an integrated, interactively expandable view of the network context of the input proteins enriched with biological information associated with these proteins.

The routine FunCoup globally reconstructs protein networks in human and other eukaryotes from comprehensive data integration, namely protein-protein interactions, mRNA expression, subcellular location, phylogenetic profiles, miRNA-mRNA targeting, transcription factor binding sites, protein expression, and domain-domain interactions [39]. The software utilizes InParanoid to transfer information between species. In the course of visualization the user is provided with the option to group networks by spatial subcellular position of proteins, their membership - relation to pathways, or as a force-directed layout. Furthermore, where possible, a detailed description of the type of association between the proteins is supported (direct physical interaction, protein complex members, metabolic reaction, regulatory/signaling).

omicsNET is another data integration framework supporting researchers throughout the process of the analysis of disease specific data in identifying and selecting potential diagnostic markers or therapeutic targets [40]. Pairwise dependencies between human

proteins are calculated based on the following data sources: gene expression profiles in normal human tissues, functional gene annotation based on gene ontologies as well as on pathway information, shared transcription factor binding site as well as miRNA profiles, information on subcellular protein localization, protein-protein interaction data, and shared protein domains. Based on these dependencies a protein network is constructed which is easily extendable and is embedded in a fully automatic downloading and importing framework capable of following the fast update cycles of scientific data repositories and data formats. Objects are centered around a general definition of biological entities based on international protein index (IPI) IDs presently covering about 68k protein sequences [41].

A: Omics repositories		
ArrayExpress	www.ebi.ac.uk/microarray-as/ae	[3]
Gene Expression Omnibus	www.ncbi.nlm.nih.gov/geo	[4]
Stanford Microarray Database	http://smd.stanford.edu	[5]
Proteomics database	www.expasy.ch/ch2d	[6]
B: Data preprocessing		
RMA	http://rmaexpress.bmbolstad.com	[9]
MAS5		[10]
dChip	http://www.dchip.org	[12]
C: Explorative analysis routines		
Bioconductor	www.bioconductor.org	[25]
SAM	http://rmaexpress.bmbolstad.com	[27]
TIGR MeV	www.tm4.org/mev.html	[28]
Functional annotation		
DAVID	http://david.abcc.ncifcrf.gov	[31]
PANTHER	www.pantherdb.org	[29]
D: Pathway analysis		
KEGG	www.genome.jp/kegg/pathway.html	[32]
PANTHER	www.pantherdb.org	[29]
KEGG spider	http://mips.helmholtz-muenchen.de/proj/keggspider	[34]
E: In-silico promoter analysis		
JASPAR	http://jaspar.genereg.net	[35]
oPOSSUM	www.cisreg.ca/cgi-bin/oPOSSUM/opossum	[36]
F: Interaction network analysis		
STRING	http://string.embl.de	[37]
FunCoup	http://funcoup.sbc.su.se	[39]

Table 1: Listing of Omics repositories, web-resources and analysis tools discussed in this chapter.

3. METHODS

In the following section the tools described above will be exemplarily applied on a publicly available gene expression dataset. Mosig and colleagues profiled the gene expression of monocytes of patients with familial hypercholesterolemia (FH) [42]. In this study microarray gene expression experiments were performed using Affymetrix HG-U133 Plus 2.0 GeneChips, each holding 54,675 unique transcripts.

3.1. Omics Data Repositories and Data Retrieval

The example dataset is deposited in the public Gene Expression Omnibus (GEO) database (www.ncbi.nlm.nih.gov/geo) hosted by NCBI reachable via the GEO accession number 'GSE6054'. The summary page of this specific record holds a short summary of the study, the experiment type, samples used in the experiment, as well as contributors. The contact details of the corresponding author as well as the date of submission are furthermore provided.

The raw data files are provided as zipped archive which includes 23 Affymetrix CEL files providing the basis for further preprocessing and analysis (see **Note 2**).

3.2. Data Preprocessing

Main data preprocessing steps involve background correction and data normalization. One tool capable of handling both tasks in a user friendly way is CARMAweb, developed at the Technical University of Graz (<https://carmaweb.genome.tugraz.at>) [43]. Creating an account in CARMAweb allows the user storing of files and results for further analysis at a later time. CARMAweb supports a number of file formats generated by the scanner software of different platforms including Affymetrix, Applied Biosystems as well as two-color systems. When using Affymetrix data the CEL files have to be uploaded to the system in order to start the preprocessing procedure as described step by step below (see **Note 3** for a detailed discussion on input parameters and resulting plots):

1. choose *New Analysis* from the tool bar
2. select *Perform an Affymetrix GeneChip analysis*
3. upload the raw data CEL files for the analysis
4. select the preprocessing method 'mas5'
5. scale the values to 200
6. check the boxes for drawing additional plots from the raw and normalized data
7. check the box *Save the normalized expression values to a text file*
8. skip the replicate handling step as there are no replicated arrays in this example data set
9. start the analysis

Id	GenBank	UniGene	Description	LocusLink	Symbol	GSM140232.CEL	GSM140233.CEL
1007_s_at	U48705	Hs.631988	discoidin domain receptor family, member 1	780	DDR1	133,1116888	129,7100459
1053_at	M87338	Hs.647062	replication factor C (activator 1) 2, 40kDa	5982	RFC2	217,6610085	239,3148494
117_at	X51757	Hs.654614	heat shock 70kDa protein 6 (HSP70B')	3310	HSPA6	781,1669739	465,5422967
121_at	X69699	Hs.469728	paired box 8	7849	PAX8	204,6355705	281,2443974
1255_g_at	L36861	Hs.92858	guanylate cyclase activator 1A (retina)	2978	GUCA1A	5,5138363	12,3289051
1294_at	L13852	Hs.16695	ubiquitin-activating enzyme E1-like	7318	UBE1L	766,9258871	742,6529846
1316_at	X55005	Hs.724	thyroid hormone receptor, alpha	7067	THRA	53,9216242	79,9992657
1320_at	X79510	Hs.437040	protein tyrosine phosphatase, non-receptor type 21	11099	PTPN21	7,5285511	6,9065730

1405_i_at	M21121	Hs.514821	chemokine (C-C motif) ligand 5	6352	CCL5	5625,3812000	5054,0114760
1431_at	J02843	Hs.12907	cytochrome P450, family 2, subfamily E, polypeptide 1	1571	CYP2E1	37,9592864	31,5216112
1438_at	X75208	Hs.2913	EPH receptor B3	2049	EPHB3	15,6320131	14,0352867
1487_at	L38487	Hs.110849	estrogen- related receptor alpha	2101	ESRRA	510,4666229	430,1333161
1494_f_at	M33318	Hs.439056	cytochrome P450, family 2, subfamily A, polypeptide 6	1548	CYP2A6	72,2235171	79,3025924
1552256_a_at	NM_005505	Hs.520348	scavenger receptor class B, member 1	949	SCARB1	294,3431947	239,6201050
1552257_a_at	NM_015140	Hs.517670	tubulin tyrosine ligase-like family, member 12	23170	TTLL12	483,2240522	425,4874463
1552258_at	NM_052871	Hs.652166	chromosome 2 open reading frame 59	112597	C2orf59	17,1870075	24,8940935
1552261_at	NM_080735	Hs.2719	WAP four- disulfide core domain 2	10406	WFDC2	35,7236355	52,4366549
1552263_at	NM_138957	Hs.431850	mitogen- activated protein kinase 1	5594	MAPK1	872,2548451	604,3554185
1552264_a_at	NM_138957	Hs.431850	mitogen- activated protein kinase 1	5594	MAPK1	676,0495879	887,2818401
1552266_at	NM_145004	Hs.521545	ADAM metallopepti dase domain 32	203102	ADAM32	40,0241198	37,3786065

Table 2: Excerpt of the file “ExpressionValues.txt” resulting from CARMAweb preprocessing.

The first six columns hold the main identifiers for all of the 54675 transcripts included on the Affymetrix HG-U133 Plus 2.0 GeneChip. The seventh column provides a short description of the gene, and the last columns hold the normalized expression values for each array (the values of the first two arrays are exemplarily shown).

Results as well as the analysis protocols are accessible after the preprocessing steps are completed. The analysis report contains a summary of the performed analysis steps as well as plots for checking the quality of given array data. The normalized expression data set that will be used for further analysis is denoted as 'ExpressionValues.txt' and can be downloaded to a local machine (**Table 2**). Features are annotated with their respective NCBI GenBank accession number, NCBI UniGene Cluster ID, NCBI Entrez Gene ID (LocusLink ID), NCBI Gene Symbol, as well as a short summary. Result files can be downloaded separately or as a compressed archive.

3.3. Identification of Differentially Expressed Genes

The preprocessed and normalized data file "ExpressionValues.txt" is the basis for the identification of differentially expressed genes (DEGs). Main interest in our study is the identification of genes that show differential expression between subjects with familial hypercholesterolemia and healthy controls. Various open-source as well as commercial tools exist for this task as outlined in the Materials section. One open-source tool that we consider very intuitive to use is the Multi Experiment Viewer (MeV) developed at The Institute for Genomic Research (TIGR) (www.tm4.org/mev.html).

MeV is perfectly capable of handling tab-delimited text files holding expression datasets such as our normalized file 'ExpressionValues.txt'. Various statistical tests are implemented in the MeV software package, among them the t-test, the Analysis of Variance (ANOVA) for multi-group comparisons, or the Statistical Analysis of Microarrays (SAM) method controlling the False Discovery Rate. The following steps result in a list of significantly differentially expressed transcripts using the SAM method (see **Note 4** for a detailed discussion on input parameters):

1. select *Load Data* from the MeV file menu
2. check *Single-color Array* in the *Expression File Loader* dialog box
3. load the file 'ExpressionValues.txt'
4. select *Significance Analysis for Microarrays* from the *Statistics* tab
5. select the *Two-class unpaired* tab

6. assign diseased samples to group A and healthy control samples to group B
7. set the number of permutations to 500
8. select *S0* using *Tusher et. al* method
9. check *no* for calculating q-values
10. select *K-nearest neighbors impute* as Imputation Engine with 10 neighbors
11. start analysis

Once the analysis has finished, the resulting SAM graph is displayed reporting the number of significantly differentially regulated genes regarding the group comparison as well as the median number of genes being false positives at a given delta threshold level (**Fig. 1**). The slider for controlling the delta value at the bottom of the graph can be used to set the false discovery rate (FDR), representing the fraction of false positive genes among the total number of all genes indicated as being differentially regulated. Usually values in the range of 5% to 10% are acceptable. In our experiment setting, a delta value of 1.156 results in 1016 significant genes and a median number of falsely significant genes of 50. Please note that these results may slightly vary due to the sequence of random permutations used in the analysis. The list of 1016 significant genes can be displayed by selecting the node *Table Views/All Significant Genes* in the folder *Analysis Results / SAM* on the left of the MeV navigation window. The table of the significant genes can be downloaded through selecting *Save cluster* from the menu.

Using the fold change criterion can further reduce the list of interesting genes to be considered for further analysis. The fold change determines how many times the expression levels for a given transcript are increased or decreased in the diseased samples as compared to the healthy individuals. Focusing on genes showing at least a two-fold change in either direction further reduces the dataset from 1016 DEGs to 97 DEGs.

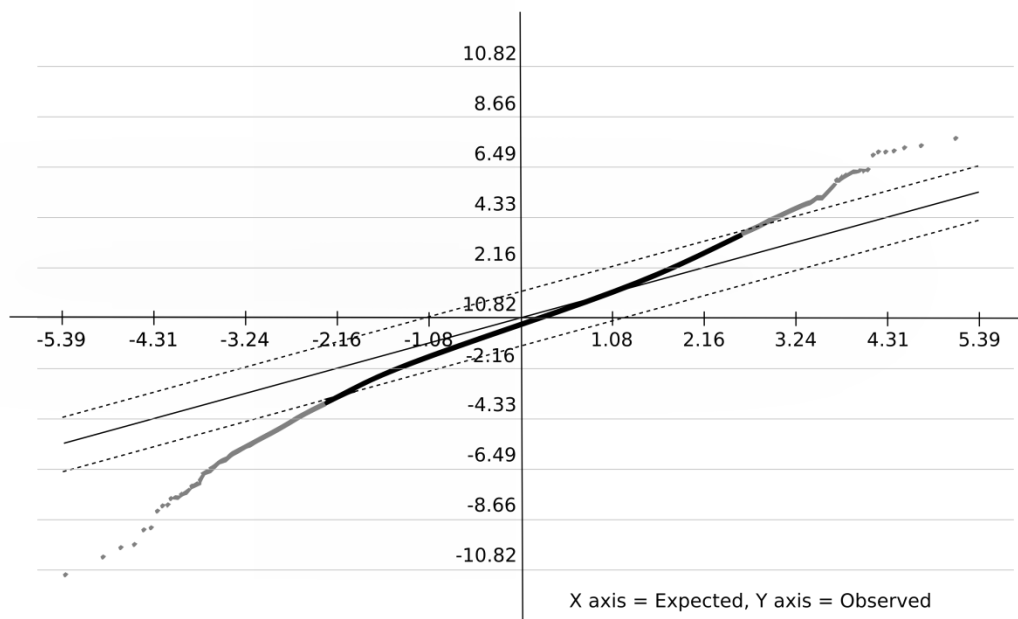


Figure 1: Example output graph resulting from a SAM analysis. The two dotted lines represent the region within \pm delta units (set to 1.156) from the observed to expected line. The genes whose plot values are within \pm delta units are considered non-significant, those above $+$ delta units are considered as significantly upregulated, and the ones below $-$ delta units are considered as significantly downregulated.

3.4. Functional Annotation and Pathway Enrichment Analysis

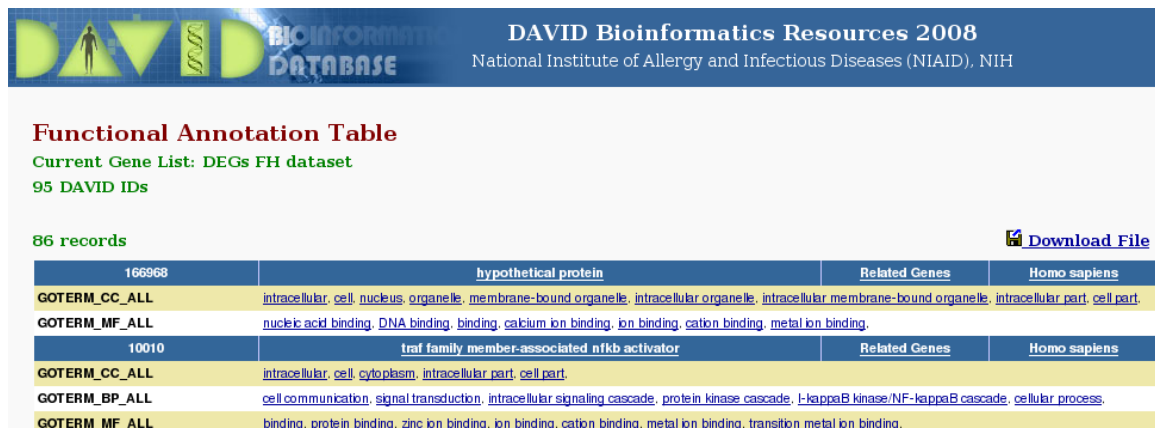
Differentially expressed genes can be linked to gene ontology categories in order to identify enriched or depleted biological processes as implemented in the DAVID tool (<http://david.abcc.ncifcrf.gov>). Input is a list of NCBI Gene IDs of e.g. differentially expressed genes or more generally speaking genes of interest that can either be pasted into the data input field provided by the application or uploaded as a simple text file. The following steps are necessary to complete the analysis:

1. select *Start Analysis* from the tool bar
2. paste the list of identifiers into 'box A' or upload the identifiers from a text file
3. select *ENTREZ_GENE_ID* as Identifier
4. select *Gene List* as List Type
5. submit list

6. choose *HOMO SAPIENS* as species in the 'List Manager'

DAVID integrates several tools for data annotation and in a first step we assign GO terms and KEGG pathways to the individual genes:

1. select *Functional Annotation Table*
2. check the boxes *GOTERM_BP_ALL*, *GOTERM_CC_ALL*, and *GOTERM_MF_ALL* from the Gene Ontology node and *KEGG_PATHWAY* from the Pathways node on the Annotation Summary Results Page
3. select *Functional Annotation Table*
4. a separate window opens showing a table with all submitted Entrez Gene IDs and their functional categories (**Fig. 2**)
5. download the table as a text file by clicking the download symbol on the upper right corner of the given window



DAVID Bioinformatics Resources 2008
National Institute of Allergy and Infectious Diseases (NIAID), NIH

Functional Annotation Table
Current Gene List: DEGs FH dataset
95 DAVID IDs

86 records [Download File](#)

Gene ID	Gene Name	Related Genes	Species
166968	hypothetical protein		Homo sapiens
GOTERM_CC_ALL	intracellular, cell, nucleus, organelle, membrane-bound organelle, intracellular organelle, intracellular membrane-bound organelle, intracellular part, cell part		
GOTERM_MF_ALL	nucleic acid binding, DNA binding, binding, calcium ion binding, ion binding, cation binding, metal ion binding		
10010	traf family member-associated nfkb activator		Homo sapiens
GOTERM_CC_ALL	intracellular, cell, cytoplasm, intracellular part, cell part		
GOTERM_BP_ALL	cell communication, signal transduction, intracellular signaling cascade, protein kinase cascade, I-kappaB kinase/NF-kappaB cascade, cellular process		
GOTERM_MF_ALL	binding, protein binding, zinc ion binding, ion binding, cation binding, metal ion binding, transition metal ion binding		

Figure 2: Example analysis output when the DAVID routine is applied. Given are the gene ontology terms for two differentially expressed genes.

Another web tool for categorizing genes by their biological function is PANTHER (www.pantherdb.org). To analyze the genes differentially expressed between FH and healthy monocytes (as given for our example case) in terms of functional enrichment when compared to the whole NCBI H. sapiens gene list, the following steps have to be performed:

1. select *Tools* from the tool bar
2. choose *Gene Expression Data Analysis and Compare gene lists*
3. select *Gene ID* as identifier and upload the list of Entrez Gene IDs for the differentially expressed genes
4. finish selecting lists
5. select *NCBI: H. sapiens genes* as reference list
6. check *Biological Processes*
7. launch analysis
8. download the results table by clicking the *Export* button on the upper left corner on the results page (**Fig. 3**)

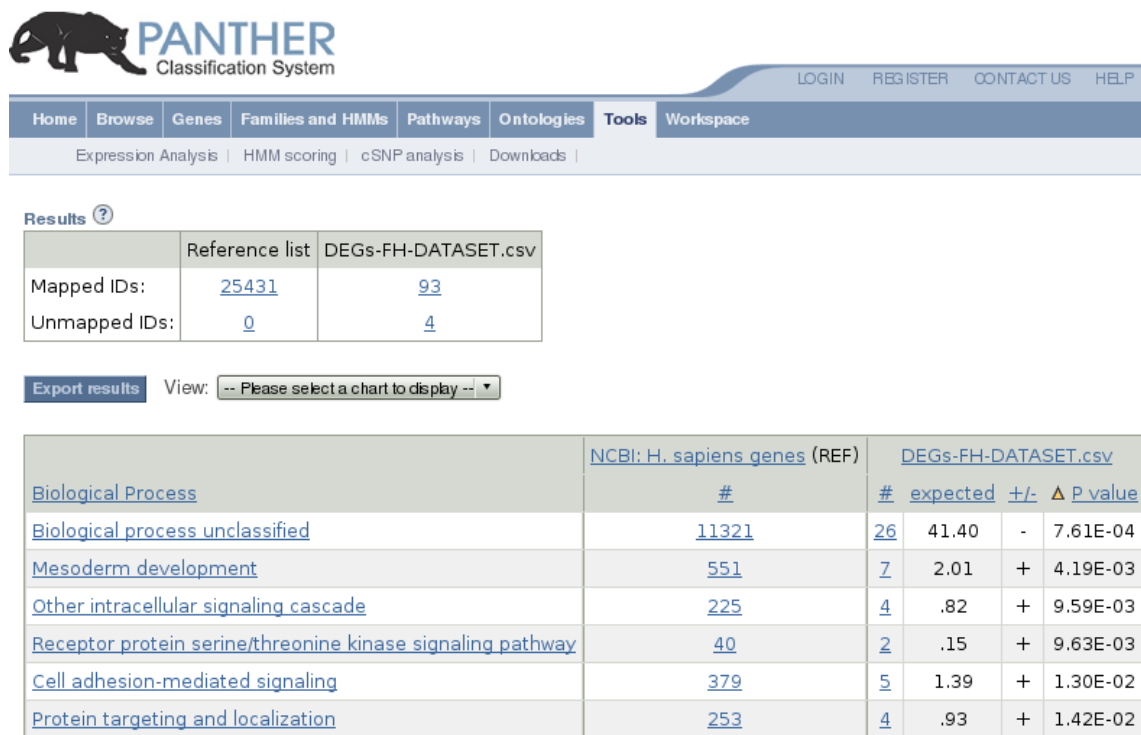


Figure 3: PANTHER analysis example output. The second and third column hold the number of genes in the reference and FH list mapping to the PANTHER classification category in the first column. The expected number of genes in the respective category is listed in column four. A plus or minus sign in the fifth column indicates over- or under-representation of features for a given category. The last column of the results table holds the p-values indicating the significance of deviation of the identified number of features with respect to the number of features present in a particular category when following a chi-square test.

3.5. In-silico Promoter Analysis

Transcription factors with enriched binding sites in a set of genes or proteins can be identified with the oPOSSUM tool (<http://www.cisreg.ca/cgi-bin/oPOSSUM/opossum>). Gene as well as protein identifiers are accepted by the analysis tool such as Ensembl IDs, HUGO Gene Symbols or aliases, RefSeq IDs, or Entrez Gene IDs.

The following steps are necessary to obtain transcription factors with enriched binding sites. For a discussion of input parameters see **Note 5**.

1. select as organism either human or mouse
2. select the type of identifier and upload your list of IDs
3. select all JASPAR Core profiles with a specificity of 10 bits
4. set the level of conservation to the top 10% of conserved regions and the matrix match threshold to 85%
5. define the region in respect to the transcription start site to be searched for binding sites
6. focus on significantly enriched transcription factors by setting the Z-score ≥ 5 and the p-value of the Fisher's exact test to ≤ 0.05

In our example the transcription factor NR2F1 is found to be significantly enriched with a p-value of < 0.001 and a Z-score of 8.069 when searching 2000 base pairs upstream of the transcription start sites of all upregulated genes. Next to the statistics the counts of transcription factor binding sites in our gene set as well as in the background gene set is given along with the transcription factor class and supergroup the transcription factor belongs to. A detailed view of the predicted binding sites in the analysis dataset is accessible via the link in the field of target gene hits.

3.6. Integrated Approaches

The STRING tool (<http://string.embl.de>) for the generation of protein interaction networks accepts both, protein identifiers or protein sequences as input. To retrieve protein identifiers from the list of differentially expressed genes, the DAVID tool can be used. The procedure is the same as described in 3.4. for the assignment of GO terms

and pathways, but the box UNIPROT_ACCESSION from the Main Accession node has to be selected. The following steps lead to a STRING network of proteins from the differentially expressed genes:

1. select the multiple names tab from the search box
2. paste the list of protein identifiers in the respective box
3. choose Homo sapiens as organism
4. start the analysis
5. review the list of input proteins and continue

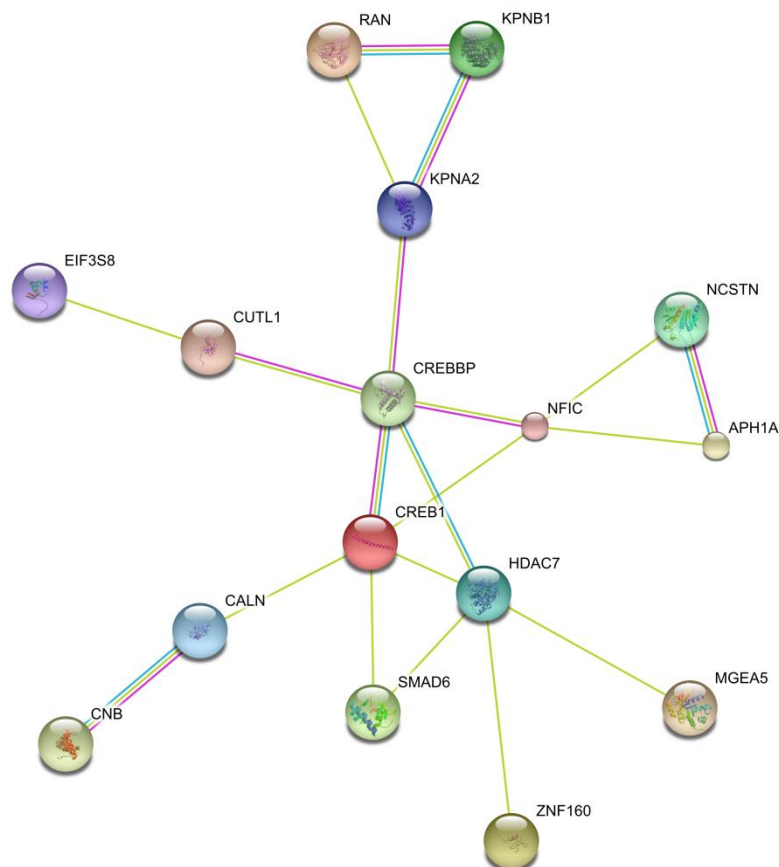


Figure 4: Subgraph extracted from the STRING protein network. Edge colors indicate the type of interaction. Olive edges: interaction based on textmining; Pink edges: experimental interaction evidence; Blue edges: information from other databases.

The resulting network holds the uploaded proteins and can be further expanded with additional interacting partners by selecting the *more* button below the graphics. The default network view is the evidence view, where nodes represent proteins and edge color indicates the type of evidence for the association. Further views can be selected on the bottom of the results page.

Fig. 4 shows a resulting subgraph when expanding the entire network of differentially expressed genes by adding ten additional partners with the highest evidence score. For the given example most of the members are involved in mRNA transcription

4. NOTES

1. A listing of databases, web-based resources and tools discussed in this work is given in **Table 1**.

2. Next to the zipped CEL files, the GEO accession summary page provides links to three additional files holding information on metadata and the normalized expression values. The SOFT formatted family file and the MINiML formatted family file include information about the family of the specific accession in text or XML format, respectively. Family implies all records related to the accession, including platform, sample, and series records. The third file is called 'Series Matrix File' and is a text file, holding expression values for all samples in matrix format. The header of these files contains all relevant metadata including the abstract, contributors, sample hybridization protocol, processing method, etc., and can be used as input for analysis software packages like the TIGR MeV tool.

3. CARMAweb provides several different methods for preprocessing, including MAS5, RMA, and additionally custom normalization can be defined. The custom normalization allows the user to select from various methods for the consecutive steps of the preprocessing procedure. In order to make arrays comparable, the expression values are scaled up or down using a pre-defined intensity value, which is by default set to 200 when using MAS5 in CARMAweb. A histogram and a boxplot of the raw data as well as of the normalized data are drawn after checking the respective box. These plots can give a first impression of the data and array quality. If a dataset includes array replicates, they can be merged by calculating the mean expression values across the replicates.

4. SAM is implemented for two-class unpaired, two-class paired, multi-class, censored survival, and one-class group comparisons. Because the FH dataset used in the given example case consists of two groups (diseased and healthy) and no pairing of samples is available, we choose the two-class unpaired design. For our dataset we consider 500 permutations to be sufficient for reaching robust results. This number however can be increased up to the point where all possible permutations are performed. If a number higher than the possible number of unique permutations is entered the user is asked whether to use all possible permutations. The S_0 constant minimizes the coefficient of variation of the relative difference in gene expression and is computed as a percentile based on alpha, which indicates the probability of false positive results. Q-

values can be computed to indicate the lowest false discovery rate at which the transcript is denoted as significant. For imputation of missing values, SAM provides two methods, namely the K-nearest neighbor algorithm and the row average method. The K-nearest neighbor algorithm replaces missing values with the k nearest neighbors according to the Euclidean distance, whereas the row average method simply uses the mean of the expression values for the respective transcript over all arrays.

5. In order to reduce the number of false positive predictions the use of more stringent input parameters is advised. We only use transcription factor binding matrices with a minimum specificity of 10 bits and a matrix match threshold of 85%. Additionally, only the top 10% of conserved regions with a minimum conservation of 70% are used.

REFERENCES

1. Wittner BS, Sgroi DC, Ryan PD, et al. Analysis of the MammaPrint breast cancer assay in a predominantly postmenopausal cohort. *Clin. Cancer Res* **2008**; 14:2988-2993.
2. Perco P, Rapberger R, Siehs C, et al. Transforming omics data into context: bioinformatics on genomics and proteomics raw data. *Electrophoresis* **2006**; 27:2659-2675.
3. Parkinson H, Kapushesky M, Shojatalab M, et al. ArrayExpress--a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res* **2007**; 35:D747-750.
4. Barrett T, Troup DB, Wilhite SE, et al. NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Res* **2009**; 37:D885-890.
5. Demeter J, Beauheim C, Gollub J, et al. The Stanford Microarray Database: implementation of new analysis tools and open source release of software. *Nucleic Acids Res* **2007**; 35:D766-770.
6. Hoogland C, Mostaguir K, Sanchez J-C, Hochstrasser DF, Appel RD. SWISS-2DPAGE, ten years later. *Proteomics* **2004**; 4:2352-2356.
7. Smolka M, Zhou H, Aebersold R. Quantitative protein profiling using two-dimensional gel electrophoresis, isotope-coded affinity tag labeling, and mass spectrometry. *Mol. Cell Proteomics* **2002**; 1:19-29.
8. Brazma A, Hingamp P, Quackenbush J, et al. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet* **2001**; 29:365-371.
9. Irizarry RA, Hobbs B, Collin F, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **2003**; 4:249-264.
10. Statistical Algorithms Reference Guide. http://www.affymetrix.com/support/technical/technotes/statistical_reference_guide.pdf. **2001**; Available at: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.97.8642>.
11. Schadt EE, Li C, Ellis B, Wong WH. Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data. *J. Cell. Biochem. Suppl* **2001**; Suppl 37:120-125.
12. Li C, Hung Wong W. Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biol* **2001**; 2:RESEARCH0032.
13. Troyanskaya O, Cantor M, Sherlock G, et al. Missing value estimation methods for DNA microarrays. *Bioinformatics* **2001**; 17:520-525.
14. Zhou X, Wang X, Dougherty ER. Missing-value estimation using linear and non-linear regression with Bayesian gene selection. *Bioinformatics* **2003**; 19:2302-2307.

15. Bø TH, Dysvik B, Jonassen I. LSImpute: accurate estimation of missing values in microarray data with least squares methods. *Nucleic Acids Res* **2004**; 32:e34.
16. Jörnsten R, Wang H-Y, Welsh WJ, Ouyang M. DNA microarray data imputation and significance analysis of differential expression. *Bioinformatics* **2005**; 21:4155-4161.
17. Grosse-Coosmann F, Boehm AM, Sickmann A. Efficient analysis and extraction of MS/MS result data from Mascot result files. *BMC Bioinformatics* **2005**; 6:290.
18. Diehn M, Sherlock G, Binkley G, et al. SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data. *Nucleic Acids Res* **2003**; 31:219-223.
19. Safran M, Chalifa-Caspi V, Shmueli O, et al. Human Gene-Centric Databases at the Weizmann Institute of Science: GeneCards, UDB, CroW 21 and HORDE. *Nucleic Acids Res* **2003**; 31:142-146.
20. Westfall PH, Young SS. Resampling-based multiple testing: examples and methods for p-value adjustment. In: *Wiley series in probability and mathematical statistics*. Wiley, 1993.
21. Dudoit S, Shaffer JP, Boldrick JC. Multiple Hypothesis Testing in Microarray Experiments. *Statist. Sci* **2003**; 18:71-103.
22. Ge Y, Dudoit S, Speed TP. Resampling-based multiple testing for microarray data analysis. *Test* **2003**; 12:1-77.
23. Nie L, Wu G, Zhang W. Statistical application and challenges in global gel-free proteomic analysis by mass spectrometry. *Crit. Rev. Biotechnol* **2008**; 28:297-307.
24. van der Laan MJ, Dudoit S, Pollard KS. Multiple testing. Part II. Step-down procedures for control of the family-wise error rate. *Stat Appl Genet Mol Biol* **2004**; 3:Article14.
25. Gentleman RC, Carey VJ, Bates DM, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* **2004**; 5:R80.
26. Efron B, Tibshirani R. *An Introduction to the Bootstrap* (Chapman & Hall/CRC Monographs on Statistics & Applied Probability). Chapman and Hall/CRC, 1994. Available at: <http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20&path=ASIN/0412042312>.
27. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. U.S.A* **2001**; 98:5116-5121.
28. Saeed AI, Sharov V, White J, et al. TM4: a free, open-source system for microarray data management and analysis. *BioTechniques* **2003**; 34:374-378.
29. Mi H, Lazareva-Ulitsky B, Loo R, et al. The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Res* **2005**; 33:D284-288.

30. Khatri P, Drăghici S. Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics* **2005**; 21:3587-3595.
31. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **2009**; 4:44-57.
32. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **2000**; 28:27-30.
33. Joshi-Tope G, Gillespie M, Vastrik I, et al. Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res* **2005**; 33:D428-432.
34. Antonov AV, Dietmann S, Mewes HW. KEGG spider: interpretation of genomics data in the context of the global gene metabolic network. *Genome Biol* **2008**; 9:R179.
35. Portales-Casamar E, Thongjuea S, Kwon AT, et al. JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res* **2010**; 38:D105-110.
36. Ho Sui SJ, Mortimer JR, Arenillas DJ, et al. oPOSSUM: identification of over-represented transcription factor binding sites in co-expressed genes. *Nucleic Acids Res* **2005**; 33:3154-3164.
37. von Mering C, Jensen LJ, Kuhn M, et al. STRING 7--recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res* **2007**; 35:D358-362.
38. Jensen LJ, Kuhn M, Stark M, et al. STRING 8--a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Research* **2009**; 37:D412-D416.
39. Alexeyenko A, Sonnhammer ELL. Global networks of functional coupling in eukaryotes from comprehensive data integration. *Genome Res* **2009**; 19:1107-1116.
40. Bernthaler A, Mühlberger I, Fechete R, Perco P, Lukas A, Mayer B. A dependency graph approach for the analysis of differential gene expression profiles. *Mol Biosyst* **2009**; 5:1720-1731.
41. Kersey PJ, Duarte J, Williams A, Karavidopoulou Y, Birney E, Apweiler R. The International Protein Index: an integrated database for proteomics experiments. *Proteomics* **2004**; 4:1985-1988.
42. Mosig S, Rennert K, Büttner P, et al. Monocytes of patients with familial hypercholesterolemia show alterations in cholesterol metabolism. *BMC Med Genomics* **2008**; 1:60.
43. Rainer J, Sanchez-Cabo F, Stocker G, Sturn A, Trajanoski Z. CARMAweb: comprehensive R- and bioconductor-based web service for microarray data analysis. *Nucleic Acids Res* **2006**; 34:W498-503.

2.1.1 The Thesis Author's Contribution

The author of the thesis designed the bioinformatics and data workflow for this methodological concept paper and conducted the specific data retrieval and analysis steps of this work.

In detail, the following contributions are due to the thesis author's efforts:

- Retrieval of the gene expression dataset on familial hypercholesterolemia from the Gene Expression Omnibus database
- Preprocessing of the dataset including background correction and raw data normalization using the CARMAweb tool
- Identification of differentially expressed genes using the SAM (Significance Analysis for Microarrays) provided by the TIGRMeV software
- Functional annotation and enrichment analysis using the DAVID tool
- Functional annotation and enrichment analysis using the PANTHER classification tool
- Converting gene identifiers to protein identifiers using the DAVID tool
- Generation of protein interaction networks according to the STRING tool
- Writing and preparation of the text describing the performed analysis steps
- Preparation of tables and figures presented in the Methods section
- Lead in preparing the manuscript draft

Biomarkers in Renal Transplantation Ischemia Reperfusion Injury

Irmgard Mühlberger¹, Paul Perco¹, Raul Fechete¹, Bernd Mayer¹, and
Rainer Oberbauer^{2,3,4*}

¹ emergentec biodevelopment GmbH, Vienna, Austria

² Department of Nephrology, Medical University of Vienna, Vienna, Austria

³ Austrian Dialysis and Transplant Registry, Wels, Austria

⁴ Krankenhaus der Elisabethinen Linz, Austria

* Corresponding author:

Rainer Oberbauer, M.D., M.Sc., F.A.S.N.,
Department of Nephrology, Medical University of Vienna
Währinger Gürtel 18 –20
A-1090 Vienna, Austria.
E-mail: rainer.oberbauer@meduniwien.ac.at

SUMMARY

Ischemia reperfusion injury (IRI) is a choreographed process leading to delayed graft function (DGF) and reduced long term patency of the transplanted organ. Early identification of recipients of grafts at risk would allow modification of the post transplant management, and thereby potentially improve short and long term outcomes.

The recently emerged 'omics' technologies together with bioinformatics work-up have allowed the integration and analysis of IRI-associated molecular profiles in the context of DGF. Such a systems biology approach promises qualitative information about interdependencies of complex processes such as IRI regulation, rather than offering descriptive tables of differentially regulated features on a transcriptome, proteome or metabolome level lacking the functional, biological framework.

In deceased donor kidney transplantation as the primary etiology resulting in IRI and DGF, a distinct signature and choreography of molecular events in the graft before harvesting appears to be associated with subsequent DGF. A systems biology assessment of these molecular changes suggests that processes along inflammation are of pivotal importance for the early stage of IRI. The causal proof of this association has been tested by a double-blinded RCT of steroid or placebo infusion into deceased donors before the organs were harvested. Thorough systems biology analysis revealed a panel of biomarkers with excellent discrimination.

In summary, integrated analysis of omics data has brought forward biomarker candidates and candidate panels which promise early assessment of IRI. The clinical utility of these markers, however, still needs to be established in prospective trials in independent patient populations.

INTRODUCTION

Renal transplantation is the treatment of choice for end stage renal disease but this option is limited by the availability of donor organs. Deceased donor transplantation accounts for the majority of transplants performed in most parts of Europe and the United States, however the short and long term outcomes are worse compared to live donor kidney transplantation. There are several explanations for this phenomenon, but certainly the autonomous cytokine storm after brain death in the donor characterized by central diabetes insipidus and subsequent hemodynamic instability heavily contributes to this incident. In comparison, kidneys from live donors almost never show signs of inflammation and acute tubular damage as evidenced from biopsies obtained after harvest but before engraftment [1]. Tubular and vascular damage in the donor organ after cold ischemia but before transplantation is highly associated with subsequent ischemic reperfusion injury (IRI) and delayed graft function (DGF). In fact according to large registries such as the UNOS/SRTR, recipients of standard criteria deceased donor organs experience DGF defined as more than one post-transplant dialysis in roughly 20% of cases. Recipients of organs from extended criteria donors, or donors with cardiac death exhibit an even higher rate of up to 50% primary non-function [2]. Graft survival of organs with DGF is dramatically impaired compared to primary functioning kidney grafts. The relative risk of graft failure is 1.5 to 2.5 fold higher in DGF compared to primary functioning grafts [2-4]. By appreciating this dramatic impact of DGF on outcomes it becomes obvious that acute renal failure is in fact not a 'cute' renal injury but a rather devastating condition which needs to be prevented by all means. Prevention requires identification of subjects at risk before the event occurred. Thus we were set out over the last decade to search for potential biomarkers for IRI and DGF in donor kidneys on a genome wide basis.

The performance characteristics and validity of such biomarkers, however, is difficult to assess since this process requires the analysis of the derived markers with morphological grading of the allograft as gold standard for tissue injury. Furthermore, thorough clinical follow up of recipients of these allografts is mandatory. And lastly, the predictive values of these biomarker tests depend on the incidence of IRI. So far no kidney biomarker exists that exhibits adequate discrimination and calibration for useful clinical application. It is likely that a panel of a few, rather than single biomarkers will be used in the future to identify subjects at risk for IRI and DGF.

If subjects at risk could be identified with adequate precision, prophylactic measures would be feasible potentially leading to a reduced rate and severity of IRI - and hopefully longer graft patency. The next paragraphs provide an overview on biomarker discovery and verification for the prediction of IRI, and their utility for clinical use.

HUMAN STUDIES

A number of studies have been performed utilizing animal models of IRI [5-7]. Supavekin and colleagues report on expression analyses using cDNA microarrays in a mouse model identifying 91 upregulated and 156 downregulated genes after ARF induction with a significant number involved in apoptotic processes [7]. Yoshida et al. identified 109 differentially expressed genes in a mouse model with ischemia reperfusion injury induced ARF [6]. In a similar setting in a rat model the same group reported 18 genes as being differentially expressed after IRI induced ARF [5]. However, a shortcoming of these animal trials is their unclear resemblance of the human situation. Comparing e.g. differential gene expression profiles from the studies mentioned above provided only partial overlap on the level of involved features [8].

The first genome wide gene expression studies in human donor kidney biopsies were performed more than five years ago. Hauser and colleagues showed that genes participating in the functional ontologies of inflammation and immune response were the primary predictors of subsequent IRI and DGF [9]. These data were confirmed on the protein level for selected candidates of these GO families such as the adhesion molecules ICAM-1, VCAM and ELAM [10]. The immunohistochemistry studies showed, as expected, varying expression of protein markers of inflammation in different compartments of the kidney such as the tubulointerstitium, the vessels and the glomerular capillary loops. Thus subsequent studies used laser capture microdissection to separate the functional units of the nephron before analyzing compartment specific differential gene expression using live donor kidney biopsies obtained immediately before transplantation as controls. Kainz and colleagues demonstrated that gene expression profiles are distinctly different not only between the compartments, i.e. glomeruli and the tubulointerstitium, but also between deceased and live donor kidneys. Again, members of the inflammation and immune response family were the main discriminators between the compartments and organ sources.

Mueller et al. also analyzed the transcriptome of zero hour donor kidney biopsies and reported a gene set consisting of 1051 transcripts differentially expressed between a group of organs from deceased donors with greater incidence of delayed graft function as compared to a group of organs from deceased donors with primary function [11]. Mas and colleagues identified 36 candidate genes associated with delayed graft function in deceased donor kidney biopsies with a large fraction being involved in inflammatory responses [12].

A summary of biomarker candidates presently discussed in the literature in the context of IRI and DGF is given in Table 1.

Gene Name	Gene Symbol	References
actin, alpha 2, smooth muscle, aorta	ACTA2	Badid et al. [13]
uromodulin	UMOD	Lynn and Marshall [14]; Zimmerhackl [15]
lectin, galactoside-binding, soluble, 3	LGALS3	Nishiyama et al. [16]
spermidine/spermine N1-acetyltransferase 1	SAT1	Zahedi et al. [17]
hepatitis A virus cellular receptor 1	HAVCR1	Ichimura et al.[18]; Hong et al. [19]; Vaidya [20]
chemokine (C-X-C motif) ligand 1	CXCL1	Molls et al.[21]
annexin A2	ANXA2	Cheng et al. [22]
S100 calcium binding protein A6	S100A6	Cheng et al. [22]
cysteine-rich, angiogenic inducer, 61	CYR61	Muramatsu et al. [23]
S100 calcium binding protein B	S100B	Pelinka et al. [24]
alpha-1-microglobulin/bikunin precursor	AMBP	Herget-Rosenthal et al. [25]
lipocalin 2	LCN2	Mishra et al. [26]
complement component 3	C3	Farrar et al. [27]
fatty acid binding protein 1, liver	FABP1	Yamamoto et al. [28]; Pelsers et al. [29]
activating transcription factor 3	ATF3	Zhou et al. [30]; Yoshida et al. [31]
Netrin 1	NTN1	Reeves et al. [32]
endoglin	ENG	Docherty et al. [33]
guanylyl cyclase G	GUCY2G	Lin et al. [34]
BH3 interacting domain death antagonist	BID	Wei et al. [35]
B-Cell CLL/lymphoma 2	BCL2	Valdes et al. [36]; Waller et al. [37]
BCL2-associated X protein	BAX	Valdes et al. [36]; Waller et al. [37]
Prostaglandin-endoperoxide synthase 2	PTGS2	Villanueva et al. [38]; Matsuyama et al. [39]
ADAM metalloproteinase with thrombospondin type 1 motif, 1	ADAMTS1	Basile et al. [40]
Cyclin-dependent kinase inhibitor	CDKN1A	Chkhotua et al. [41]; Hochegger et al. [42]

Table 1: Biomarker candidates in the context of IRI and DGF as reported in the literature. Provided is the gene name, the gene symbol, and the respective scientific references.

Most of the omics studies performed so far reported features differentially regulated on the transcriptome or proteome level. However, such descriptive lists are hardly amenable for a functional interpretation with respect to associated processes and pathways. For addressing this issue, subsequent approaches in that very field were designed to enhance the understanding of the choreographed processes by using extended bioinformatics [44]. Systems biology is one means where information characterizing IRI on different cellular layers as genome wide gene expression or proteomics are incorporated in the data analysis to better identify functionally interlinked molecular processes (instead of descriptive feature lists), and on this basis an improved identification of biomarkers which potentially predict biological events such as IRI and DGF [45,46] might become feasible.

Effects of IRI on medium term graft function as well as other related outcomes such as ESA use in the first year after engraftment were recently studied by such approaches [47,48]. Perco and colleagues, as well as Wilflingseder and coworkers identified molecular predictors in the donor kidney biopsy supporting the prediction of the graft status one year after implantation. The accuracy of this approach provided an explanation of 28% of the variability of one year serum creatinine using a biomarker panel, whereas morphological criteria (CADI score) together with clinical variables performed much poorer (adjusted R^2 of 14%). The main predictors came from the NLR protein family, pyrin domain containing 2 (NLRP2), immunoglobulin J polypeptide (IGJ), and the regulator of G-protein signaling 5 (RGS5), again indicating the central role of immune response signaling [47].

Similarly, the use of ESA requirement in the first year after engraftment is more prevalent in subjects who experience IRI and subsequent DGF. Wilflingseder and colleagues [49] found that regulators of immunity and inflammation may be used as biomarkers for IRI and subsequent ESA dependency even when adjusted for variables known to be associated with anemia including donor age, biopsy confirmed acute

rejection, serum CRP levels or GFR. The AUC of the ROC curve for the prediction of ESA dependency was 0.93 in the molecular predictor model but only 0.84 in the model of clinical variables [49]. The authors found three specific genes SPRR2C (small proline-rich protein 2C pseudogene), B3GALTL (beta-1,3-galactosyltransferase-like), and GSTT1 (glutathione S-transferase theta 1), which are now further evaluated as biomarkers for ESA dependency.

The usefulness of the information in terms of biomarker utility has certainly improved over the last years by providing qualitative information on IRI and DGF associated molecular processes. Nevertheless, the assessment accuracy on the basis of the presently given biomarker spectrum is still rather poor. This finding might be grounded on the considerable false positive rate of omics results, partially based on experimental heterogeneity as well as on shortcomings of applied statistical analysis procedures. Therefore we set out to incorporate given IRI associated omics profiles in a fully integrated systems biology framework.

THE ‘omicsNET’ DATA INTEGRATION APPROACH

Hauser and colleagues performed a transcriptomics study comparing live and diseased kidney donor organs, and identified 90 genes as differentially regulated [9]. As outlined above, the incidence of postischemic acute transplant kidney failure is significantly increased when implanting donor organs from deceased subjects. The main functional roles of the corresponding genes according to the PANTHER (Protein ANalysis THrough Evolutionary Relationship) Classification System (<http://www.pantherdb.org>) were immunity and defense as well as metabolism, as presented in Table 2A. Significant categories were identified using chi-square test statistics of assigned genes as compared to a reference gene set of all assigned human genes.

We further analyzed this data set in a computational systems biology framework following an interaction network analysis: The methodological basis of this approach is computational delineation of dependencies between human genes and proteins which are derived by inclusion of a broad omics data spectrum: Each gene/protein is represented as object (node in the interaction network) and characterized by associated functional annotation terms (stemming from e.g. gene ontologies), the given

genes' reference expression as determined for 32 tissues, experimentally derived interaction data of encoded proteins, as well as the proteins' subcellular location. On the basis of this object annotation we computed pair-wise object-object dependencies (representing edges between the nodes) applying a functional utilizing the annotation data as parameters. The resulting reference graph therefore encodes an estimate on the (functional) dependencies between genes and proteins. We then mapped the 90 features found to be differentially regulated between live and diseased donor organs on the corresponding gene objects of the reference graph, and computed the shortest paths between these nodes. The resulting subgraph is given in Figure 1.

The subgraph (holding in total 84 gene/protein nodes) derived on the basis of the gene expression profile holds all genes being statistically significantly differentially expressed (blue), includes all genes/proteins interconnecting the expression profile representatives (grey), and identifies the interconnecting nodes belonging to the functional category inflammation, given in orange. Obvious is the significant enrichment of inflammation-associated genes encoded in this subgraph, as also found when computing significantly enriched biological processes as provided in Table 2B.

biological process	p-value
(A) statistical analysis	
Immunity and defense	1.8E-04
Proteolysis	1.1E-04
Lipid metabolism	9.4E-03
Amino acid metabolism	1.2E-02
Complement-mediated immunity	1.3E-02
(B) dependency graph analysis	
Immunity and defense	3.2E-32
Signal transduction	6.0E-29
Cell proliferation	7.5E-28
Blood clotting	2.6E-18
Protein phosphorylation	4.3E-18

Table 2: PANTHER biological processes and their significance of population expressed as p-value following a chi-square test for **(A)** statistical analysis of the gene expression data alone and **(B)** analysis of the 84 nodes (107 edges) as defined in the context of the dependency graph (Figure 1).

As mentioned above, several studies identified markers of inflammation in the donor kidney as being associated with DGF but a causal proof was never done as this would require testing the suppression of inflammation in the donor. Two nodes of the subgraph are of particular interest, namely NFKB1 and NR3C1, as these are targets of corticosteroids. Supported by this analysis we designed a double blinded RCT to test the hypothesis whether suppression of inflammation in the donor would ameliorate IRI and subsequently reduce the rate of DGF. This study is presently ongoing.

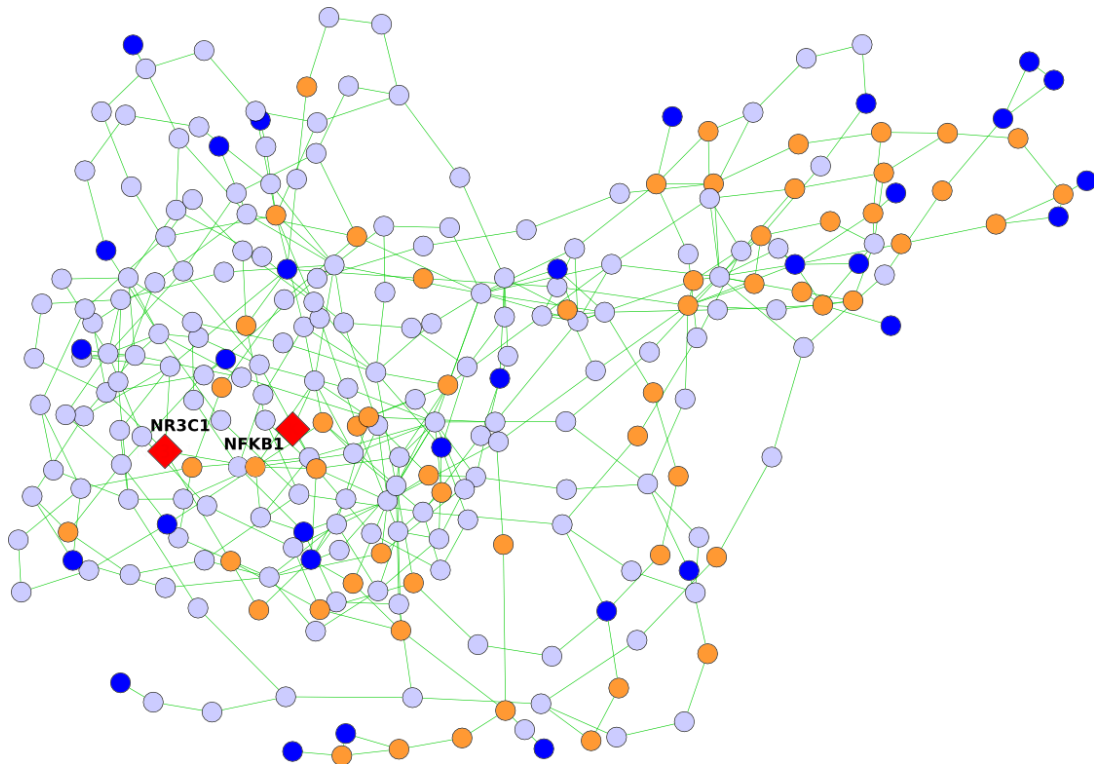


Figure 1: Subgraph interlinking significantly differentially regulated genes characterizing DGF as determined by analyzing biopsy samples of live and diseased donor kidneys. Nodes represent genes/proteins, and edges indicate dependencies between nodes. Blue colored nodes indicate genes identified as significantly differentially regulated, grey colored nodes resemble members of shortest paths connecting experimentally identified nodes, and orange colored nodes being members of the functional category inflammation. Red squares indicate the corticosteroid targets NR3C1 and NFKB1.

We repeated this dependency graph analysis approach utilizing the biomarker candidate list given in Table 1, and the results are shown in Table 3.

biological process	p-value
(A) statistical analysis	
Angiogenesis	4.19E-05
Oncogenesis	1.66E-04
Mesoderm development	1.86E-03
Developmental processes	4.00E-02
Cell proliferation and differentiation	1.23E-02
(B) dependency graph analysis	
Cell proliferation and differentiation	3.31E-15
Signal transduction	6.79E-14
Oncogenesis	8.04E-11
Immunity and defense	5.59E-10
Protein phosphorylation	6.74E-10

Table 3: PANTHER biological processes and their significance of population expressed as p-value following a chi-square test for **(A)** statistical analysis of the candidate biomarkers alone and **(B)** analysis of the 262 nodes (411 edges) found when identifying the candidate biomarker-associated dependency subgraph.

On the basis of the public domain candidate list, inflammation events as represented by immunity and defense mechanisms is not over-represented, as no respective entries are given in Table 3A. Top ranked category is angiogenesis with the majority of genes being anti-angiogenic. This is interesting since hypoxia following ischemia reperfusion injury would suggest an upregulation of pro-angiogenic factors. Basile and colleagues investigated angiogenesis related genes in the context of ischemia reperfusion injury in more detail in a microarray study and identified ADAMTS1, a secreted VEGF inhibitor, as being highly upregulated after IRI [40]. Rudnicki et al. deciphered the connection between the VEGF pathway and hypoxia in the setting of chronic kidney disease, and found a decreased expression of VEGF although hypoxia inducible factors were highly upregulated in patients with a progressive course of disease [50]. Other processes identified involve developmental events and apoptosis.

After mapping the 25 biomarker candidates on the dependency reference graph and determining the shortest paths interlinking the candidates, 237 additional nodes were

found in the subgraph. Computing the biological processes over-represented on this level ranks cell proliferation and differentiation as most prominent process, and immunity and defense also emerges as being relevant categories (Table 3B).

CLINICAL APPLICATION

A number of studies have been carried out recently to evaluate the role of the most promising biomarkers in the prediction of clinical outcomes in acute renal failure (ARF). Liangos et al. conducted a cohort study of 198 hospitalized patients with ARF, 37% among them with an ischemic cause. They showed that HAVCR1 (KIM-1) levels are associated with measurements of disease severity [51].

However, the clinical utility of markers as KIM1 and NGAL in the context of DGF still has to be demonstrated. We have started a clinical study for evaluating the impact of inflammation on DGF. This clinical study includes steroid pretreatment of the deceased organ donor to detect a reduction in the rate of IRI and DGF from the current 25% to 12.5% with adequate power (Current Controlled Trials Registration ISRCTN78828338). The trial with a calculated sample size of 420 required a protocol donor kidney biopsy which was subjected to genomics and Systems Biology analyses. As of the end of 2008 all subjects have been enrolled and most of the analysis performed. The interim analysis after half of the sample size has been enrolled was presented at the annual American Society of Nephrology meeting in 2008 [52]. This analysis showed nice suppression of inflammation in the steroid group and perfect discrimination by treatment. The effect on the clinical endpoint DGF however was not affected in a way that would have allowed stopping the trial by the pre-specified alpha error boundaries which were set according to the Lan DeMets extension of the O'Brian Fleming criteria. The analysis of the full data set will be presented in a full paper in 2009, then providing more evidence on the potential role of inflammation in IRI and DGF.

Besides the causal proof of molecular features which are associated with early graft dysfunction, validation of candidate biomarkers in an independent sample and different types of accessible body fluids such as blood or urine is required. These requests suggest that the identified molecular feature on the mRNA level cause also a differential abundance of respective protein(s). These protein(s) need to be secreted from injured renal cells of any compartment whatsoever and have to have certain kinetics, i.e. a half life of few hours which would allow detection of dynamic changes.

Finally, even before the test characteristics can be checked, a reliable measurement of the concentration of this(ese) protein(s) need to be established e.g. utilizing ELISA other approaches.

CONCLUSION

Novel biomarker candidates for assessment of IRI emerged recently, and omics techniques have provided a major contribution to these discoveries. However, so far there is no clear 'best' predictive marker that has been validated in independent samples neither on mRNA nor on protein level. Novel analysis procedures as systems biology approaches might provide further insight into the cellular processes characterizing IRI, which in turn will allow selection of superior biomarker candidates. Following the present data status inflammation events may be early stage indicators of IRI, triggering subsequent events along cell proliferation and apoptosis.

REFERENCES

1. Kainz A, Mitterbauer C, Hauser P, et al. Alterations in gene expression in cadaveric vs. live donor kidneys suggest impaired tubular counterbalance of oxidative stress at implantation. *Am. J. Transplant* **2004**; 4:1595-1604.
2. UNOS/SRTR. Annual Report of the U.S. Organ Procurement and Transplantation Network and the Scientific Registry of Transplant Recipients: Transplant Data 1997-2006 Health Resources and Services Administration. Rockville: Healthcare Systems Bureau, Division of Transplantation, **2007**.
3. Ojo AO, Wolfe RA, Held PJ, Port FK, Schumouder RL. Delayed graft function: risk factors and implications for renal allograft survival. *Transplantation* **1997**; 63:968-974.
4. Heinze G, Collins S, Benedict MA, et al. The association between angiotensin converting enzyme inhibitor or angiotensin receptor blocker use during postischemic acute transplant failure and renal allograft survival. *Transplantation* **2006**; 82:1441-1448.
5. Yoshida T, Tang S-S, Hsiao L-L, Jensen RV, Ingelfinger JR, Gullans SR. Global analysis of gene expression in renal ischemia-reperfusion in the mouse. *Biochem. Biophys. Res. Commun* **2002**; 291:787-794.
6. Yoshida T, Kurella M, Beato F, et al. Monitoring changes in gene expression in renal ischemia-reperfusion in the rat. *Kidney Int* **2002**; 61:1646-1654.
7. Supavekin S, Zhang W, Kucherlapati R, Kaskel FJ, Moore LC, Devarajan P. Differential gene expression following early renal ischemia/reperfusion. *Kidney Int* **2003**; 63:1714-1724.
8. Perco P, Pleban C, Kainz A, Lukas A, Mayer B, Oberbauer R. Gene expression and biomarkers in renal transplant ischemia reperfusion injury. *Transpl. Int* **2007**; 20:2-11.
9. Hauser P, Schwarz C, Mitterbauer C, et al. Genome-wide gene-expression patterns of donor kidney biopsies distinguish primary allograft function. *Lab. Invest* **2004**; 84:353-361.
10. Schwarz C, Regele H, Steininger R, Hansmann C, Mayer G, Oberbauer R. The contribution of adhesion molecule expression in donor kidney biopsies to early allograft dysfunction. *Transplantation* **2001**; 71:1666-1670.
11. Mueller TF, Reeve J, Jhangri GS, et al. The transcriptome of the implant biopsy identifies donor kidneys at increased risk of delayed graft function. *Am. J. Transplant* **2008**; 8:78-85.
12. Mas VR, Archer KJ, Yanek K, et al. Gene expression patterns in deceased donor kidneys developing delayed graft function after kidney transplantation. *Transplantation* **2008**; 85:626-635.
13. Badid C, Desmouliere A, Babici D, et al. Interstitial expression of alpha-SMA: an early marker of chronic renal allograft dysfunction. *Nephrol. Dial. Transplant* **2002**; 17:1993-1998.

14. Lynn KL, Marshall RD. Excretion of Tamm-Horsfall glycoprotein in renal disease. *Clin. Nephrol* **1984**; 22:253-257.
15. Zimmerhackl LB. Evaluation of nephrotoxicity with renal antigens in children: role of Tamm-Horsfall protein. *Eur. J. Clin. Pharmacol* **1993**; 44 Suppl 1:S39-42.
16. Nishiyama J, Kobayashi S, Ishida A, et al. Up-regulation of galectin-3 in acute renal failure of the rat. *Am. J. Pathol* **2000**; 157:815-823.
17. Zahedi K, Wang Z, Barone S, et al. Expression of SSAT, a novel biomarker of tubular cell damage, increases in kidney ischemia-reperfusion injury. *Am. J. Physiol. Renal Physiol* **2003**; 284:F1046-1055.
18. Ichimura T, Hung CC, Yang SA, Stevens JL, Bonventre JV. Kidney injury molecule-1: a tissue and urinary biomarker for nephrotoxicant-induced renal injury. *Am. J. Physiol. Renal Physiol* **2004**; 286:F552-563.
19. Hong ME, Hong JC, Stepkowski S, Kahan BD. Correlation between cyclosporine-induced nephrotoxicity in reduced nephron mass and expression of kidney injury molecule-1 and aquaporin-2 gene. *Transplant. Proc* **2005**; 37:4254-4258.
20. Vaidya VS, Ramirez V, Ichimura T, Bobadilla NA, Bonventre JV. Urinary kidney injury molecule-1: a sensitive quantitative biomarker for early detection of kidney tubular injury. *Am. J. Physiol. Renal Physiol* **2006**; 290:F517-529.
21. Molls RR, Savransky V, Liu M, et al. Keratinocyte-derived chemokine is an early biomarker of ischemic acute kidney injury. *Am. J. Physiol. Renal Physiol* **2006**; 290:F1187-1193.
22. Cheng C-W, Rifai A, Ka S-M, et al. Calcium-binding proteins annexin A2 and S100A6 are sensors of tubular injury and recovery in acute renal failure. *Kidney Int* **2005**; 68:2694-2703.
23. Muramatsu Y, Tsujie M, Kohda Y, et al. Early detection of cysteine rich protein 61 (CYR61, CCN1) in urine following renal ischemic reperfusion injury. *Kidney Int* **2002**; 62:1601-1610.
24. Pelinka LE, Harada N, Szalay L, Jafarmadar M, Redl H, Bahrami S. Release of S100B differs during ischemia and reperfusion of the liver, the gut, and the kidney in rats. *Shock* **2004**; 21:72-76.
25. Herget-Rosenthal S, Marggraf G, Hüsing J, et al. Early detection of acute renal failure by serum cystatin C. *Kidney Int* **2004**; 66:1115-1122.
26. Mishra J, Dent C, Tarabishi R, et al. Neutrophil gelatinase-associated lipocalin (NGAL) as a biomarker for acute renal injury after cardiac surgery. *Lancet* **2005**; 365:1231-1238.
27. Farrar CA, Zhou W, Lin T, Sacks SH. Local extravascular pool of C3 is a determinant of postischemic acute renal failure. *FASEB J* **2006**; 20:217-226.
28. Yamamoto T, Noiri E, Ono Y, et al. Renal L-type fatty acid-binding protein in acute ischemic injury. *J. Am. Soc. Nephrol* **2007**; 18:2894-2902.
29. Pelsers MMAL. Fatty acid-binding protein as marker for renal injury. *Scand. J. Clin. Lab. Invest. Suppl* **2008**; 241:73-77.

30. Zhou H, Cheruvanky A, Hu X, et al. Urinary exosomal transcription factors, a new class of biomarkers for renal disease. *Kidney Int* **2008**; 74:613-621.
31. Yoshida T, Sugiura H, Mitobe M, et al. ATF3 protects against renal ischemia-reperfusion injury. *J. Am. Soc. Nephrol* **2008**; 19:217-224.
32. Reeves WB, Kwon O, Ramesh G. Netrin-1 and kidney injury. II. Netrin-1 is an early biomarker of acute kidney injury. *Am. J. Physiol. Renal Physiol* **2008**; 294:F731-738.
33. Docherty NG, López-Novoa JM, Arevalo M, et al. Endoglin regulates renal ischaemia-reperfusion injury. *Nephrol. Dial. Transplant* **2006**; 21:2106-2119.
34. Lin H, Cheng C-F, Hou H-H, et al. Disruption of guanylyl cyclase-G protects against acute renal injury. *J. Am. Soc. Nephrol* **2008**; 19:339-348.
35. Wei Q, Yin X-M, Wang M-H, Dong Z. Bid deficiency ameliorates ischemic renal failure and delays animal death in C57BL/6 mice. *Am. J. Physiol. Renal Physiol* **2006**; 290:F35-42.
36. Valdés F, Pásaro E, Díaz I, et al. Segmental heterogeneity in Bcl-2, Bcl-xL and Bax expression in rat tubular epithelium after ischemia-reperfusion. *Nephrology (Carlton)* **2008**; 13:294-301.
37. Waller HL, Harper SJF, Hosgood SA, et al. Differential expression of cytoprotective and apoptotic genes in an ischaemia-reperfusion isolated organ perfusion model of the transplanted kidney. *Transpl. Int* **2007**; 20:625-631.
38. Villanueva S, Céspedes C, González AA, Vio CP, Velarde V. Effect of ischemic acute renal damage on the expression of COX-2 and oxidative stress-related elements in rat kidney. *Am. J. Physiol. Renal Physiol* **2007**; 292:F1364-1371.
39. Matsuyama M, Yoshimura R, Hase T, Kawahito Y, Sano H, Nakatani T. Study of cyclooxygenase-2 in renal ischemia-reperfusion injury. *Transplant. Proc* **2005**; 37:370-372.
40. Basile DP, Fredrich K, Chelladurai B, Leonard EC, Parrish AR. Renal ischemia reperfusion inhibits VEGF expression and induces ADAMTS-1, a novel VEGF inhibitor. *Am. J. Physiol. Renal Physiol* **2008**; 294:F928-936.
41. Chkhotua AB, Abendroth D, Froeba G, Schelzig H. Up-regulation of cell cycle regulatory genes after renal ischemia/reperfusion: differential expression of p16(INK4a), p21(WAF1/CIP1) and p27(Kip1) cyclin-dependent kinase inhibitor genes depending on reperfusion time. *Transpl. Int* **2006**; 19:72-77.
42. Hochegger K, Koppelstaetter C, Tagwerker A, et al. p21 and mTERT are novel markers for determining different ischemic time periods in renal ischemia-reperfusion injury. *Am. J. Physiol. Renal Physiol* **2007**; 292:F762-768.
43. Bellini MH, Coutinho EL, Filgueiras TC, Maciel TT, Schor N. Endostatin expression in the murine model of ischaemia/reperfusion-induced acute renal failure. *Nephrology (Carlton)* **2007**; 12:459-465.
44. Perco P, Kainz A, Mayer G, Lukas A, Oberbauer R, Mayer B. Detection of coregulation in differential gene expression profiles. *BioSystems* **2005**; 82:235-247.

45. Perco P, Rapberger R, Siehs C, et al. Transforming omics data into context: bioinformatics on genomics and proteomics raw data. *Electrophoresis* **2006**; 27:2659-2675.
46. Perco P, Wilflingseder J, Bernthaler A, 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.
47. Perco P, Kainz A, Wilflingseder J, Soleiman A, Mayer B, Oberbauer R. Histogenomics: association of gene expression patterns with histological parameters in kidney biopsies. *Transplantation* **2009**; 87:290-295.
48. Wilflingseder J, Perco P, Kainz A, Korbély R, Mayer B, Oberbauer R. Biocompatibility of haemodialysis membranes determined by gene expression of human leucocytes: a crossover study. *Eur. J. Clin. Invest* **2008**; 38:918-924.
49. Wilflingseder J, Kainz A, Perco P, Korbély R, Mayer B, Oberbauer R. Molecular predictors for anaemia after kidney transplantation. *Nephrol. Dial. Transplant* **2009**; 24:1015-1023.
50. Rudnicki M, Perco P, Enrich J, 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.
51. Liangos O, Perianayagam MC, Vaidya VS, et al. Urinary N-acetyl-beta-(D)-glucosaminidase activity and kidney injury molecule-1 level are associated with adverse outcomes in acute renal failure. *J. Am. Soc. Nephrol* **2007**; 18:904-912.
52. Wilflingseder J. A Multicenter RCT of Deceased Organ Donor Pre-Treatment with Corticosteroids for the Prevention of Postischemic Acute Renal Failure. **2007**;

2.2.1 The Thesis Author's Contribution

The thesis author contributed to the study design. Moreover, the author performed the literature research on biomarkers associated with Ischemia Reperfusion Injury and Delayed Graft Function, as well as the functional analyses steps for the literature and transcriptomics datasets. Interpretation and discussion of the results were carried out in collaboration between all authors of the publication.

In detail, the following contributions are due to the thesis author's efforts:

- Review and analysis of the selection of keywords used for the literature research, as well as selection of bioinformatics analyses tools
- Performance of the literature research in PubMed
- Functional annotation and enrichment analyses of the literature derived biomarker candidates and differentially expressed genes in deceased donor kidneys (obtained from Hauser et al. [9]) using the PANTHER classification tool
- Functional annotation and enrichment analyses of those genes that derived from the interaction network analyses using the PANTHER classification tool
- Contributions to the discussion of biomarker candidates and their clinical applications
- Lead in preparing the manuscript draft

2.3 Impaired metabolism in donor kidney grafts after steroid pretreatment. Transpl Int. 2010

Impaired metabolism in donor kidney grafts after steroid pretreatment

Wilflingseder Julia^{1,2}, Kainz Alexander^{1,2}, Mühlberger Irmgard³, Perco Paul³, Robert Langer⁴, Ivan Kristo⁵, Mayer Bernd³, and Oberbauer Rainer^{1,2*}

¹ Department of Nephrology KH Elisabethinen, Linz

² Department of Nephrology Medical University of Vienna, Vienna

³ emergentec biodevelopment GmbH, Vienna

⁴ Department of Transplantation and Surgery, Semmelweis University, Budapest, Hungary

⁵ Department of Transplant Surgery Medical University of Vienna, Vienna

* Corresponding author:

Rainer Oberbauer, M.D., M.Sc., F.A.S.N.,
Department of Nephrology, Medical University of Vienna
Währinger Gürtel 18 –20
A-1090 Vienna, Austria.
E-mail: rainer.oberbauer@meduniwien.ac.at

Published in: Transpl Int. 2010 Aug;23(8):796-804.

ABSTRACT

Background

We recently showed in a randomized control trial that steroid pre-treatment of the deceased organ donor suppressed inflammation in the transplant organ but did not reduce the rate or duration of delayed graft function (DGF). The present study sought to elucidate what factors caused DGF in the steroid treated subjects.

Methods

Genome-wide gene expression profiles were used from twenty steroid pre-treated donor organs and were analyzed on the level of regulatory protein-protein interaction networks.

Results

Significance analysis of microarrays yielded 63 significantly down-regulated sequences associated with DGF that could be functionally categorized according to PANTHER ontologies into two main biological processes: transport ($p < 0.001$) and metabolism ($p < 0.001$). The identified genes suggest hypoxia as cause of DGF which cannot be counterbalanced by steroid treatment.

Conclusions

Our data showed that molecular pathways affected by ischemia such as transport and metabolism are associated with DGF. Potential interventional targeted therapy based on these findings includes PPAR-agonists or caspase inhibitors.

KEYWORDS: bioinformatics, delayed graft function, renal transplantation, system biology, transcriptome

INTRODUCTION

Kidney transplantation is the preferred treatment of end stage renal disease because it is considerably cheaper than dialysis and allows for an almost normal life. One of the main reasons of graft failure is delayed graft function (DGF), a form of acute renal failure resulting in post-transplantation oliguria, increased allograft immunogenicity and risk of acute rejection episodes, and decreased long-term survival [1]. Roughly one third of transplant patients receiving an organ from a deceased donor develop DGF and have to be treated by dialysis until the engrafted organ resumes function. The hazard ratio for graft failure is almost twice as high in recipients who experienced DGF compared to those without initial complications [2]. Factors which contribute to DGF can be divided into donor-related and recipient-related factors. Donor-related factors include donor age, diseases such as hypertension, brain death associated causes such as hemodynamic instability, massive cytokine release and vasopressor use. A thorough discussion of donor and recipient factors contributing to DGF was published by Schwarz et al [3]. The fact that DGF is a rare exception in live kidney transplantation suggests that donor factors rather than the transplant procedure itself mainly contribute to DGF.

Next to the histopathological examination of renal biopsies the determination of gene expression profiles in donor organs poses an option to determine graft quality and even predict transplant outcome to a certain extent [4,5]. In a recent study from our group we reported a number of differentially regulated genes when comparing donor organs from living and deceased donor organs. Upregulated genes in tissue samples from deceased donors were mainly involved in inflammatory processes, complement activation, apoptosis and cell adhesion [6].

Based on these findings we initiated a randomized, double blinded, placebo controlled trial to elucidate whether pretreatment of deceased organ donors with corticosteroids (1g methylprednisolone) before organ retrieval will reduce inflammation and subsequently the rate of DGF after engraftment. One main finding of this study with 447 renal allograft recipients was that steroid pretreatment caused a reduction of inflammatory signatures in the donor kidney as monitored on the level of gene expression profiles. However neither rate nor the duration of delayed graft function was different in the treatment and placebo group. We therefore hypothesize that additional pathways next to inflammation are involved in the development of DGF. Thus the analysis of the steroid treatment arm provides a unique opportunity to investigate molecular mechanisms other than inflammation which contribute to DGF.

Brain death is associated with rapid swings in blood pressure, hypo- and hypertension, coagulopathies, pulmonary changes, hypothermia and electrolyte aberrations [7-9]. Therefore donor brain death does not only result in increased inflammation but also leads to hypoperfusion and hypoxia of the donor organ [10].

The main objective of the present work was to elucidate molecular causes of DGF that were not abolished by the steroid donor pretreatment. Specifically we compared the molecular signature of kidney biopsies from steroid treated donors with primary graft function to kidneys with DGF. We sought to identify potential new targets for intervention which ultimately may reduce the current high rate of DGF.

MATERIAL AND METHODS

Donor and recipient characteristics:

Out of the 238 recipients of steroid pretreated donor kidneys we randomly identified ten of 52 who developed DGF and matched an equal number of primary graft kidneys. Matching variables of controls were cause of donor death (stroke vs trauma) and calliper matching of donors' last creatinine and donor age.

The rationale behind the sample size was that based on previous data twenty biopsies would be sufficient to detect a more than twofold difference in the expression of 30 predefined genes at an adjusted p-value of <0.05 using the Bonferroni Holm method [6,11].

Trial design

Details on the multicenter trial may be found elsewhere (<http://www.controlled-trials.com/ISRCTN78828338> and Kainz & Wilflingseder et al. [12]. In brief 269 donors stratified for age were equally randomized in blocks of four to 1000mg of corticosteroid or placebo injection six hours before organ recovery. Before transplantation, kidney wedge biopsies were obtained and subjected to genomics analyses. The posttransplant clinical course was monitored.

The study protocol was approved by the Institutional Review Board (Ethical Committee of the Medical University of Vienna # EK-067/2005, to be found at <http://ohrp.cit.nih.gov/search/asearch.asp>) and the EUROTRANSPLANT kidney advisory committee (#6021KAC06) at each study site and conducted according to IRB standards at each institution. DGF was defined as the need for more than one dialysis treatment within the first week after transplantation or creatinine values above 3mg/dl during the first week after transplantation.

Laboratory procedures and biostatistical analyses

Donor kidney biopsy specimen, RNA isolation and amplification

All organs were perfused with a histidine-tryptophan-ketoglutarat cold preservation solution at 4°C during organ procurement [13]. The cold ischemic time was not longer than 24 hours. Wedge biopsies of each kidney were taken under sterile conditions at the end of the cold ischemic time right before transplantation. The biopsy specimens were immediately submerged in RNAlater[®] (Ambion, Austin, Texas) and stored at 4°C.

Total RNA was isolated and purified using chloroform and trizol reagent (Invitrogen, Carlsbad, California). RNA yield and quality was checked with the Agilent 2100 Bioanalyzer and RNA6000 LabChip[®] kit (Agilent, Palo Alto, California). Stratagene Universal human reference RNA was used as reference (Stratagene, La Jolla, California).

Two micrograms of isolated total RNA were amplified using the RiboAmp RNA amplification kit (Arcturus, Mountain View, California). The amplified RNA was inspected on an ethidium bromide stained 1% agarose gel and on the Agilent 2100 Bioanalyzer. For the twenty zero-hour kidney biopsies the RNA was of sufficient quality to proceed with microarray analysis.

Microarray hybridization and scanning

cDNA microarrays holding 41,421 (batch: SHEO) features were obtained from the Stanford University Functional Genomics core facility. All microarray experiments were performed as described earlier [14]. The detailed protocols are available at <http://genome-www.stanford.edu/>. Using a type II experimental setup, 1 µg of sample and standard Stratagene Universal human reference aRNA were labeled with CyScribe

cDNA post labeling kit (Amersham Pharmacia Biotech, Buckinghamshire, UK) in a two-step procedure.

Samples were loaded onto arrays and incubated for 18 hours in a 65°C water bath. After three washing steps, the fluorescence images of the hybridized microarrays were examined using a GenePix 4100A scanner (Axon Instruments, Union City, California). The GenePix Pro 6.0 software was used to grid images and to calculate spot intensities. Arrays were numbered according to the anonymous organ donor ID, and were processed in random order. Image-, grid- and data-files were submitted to the Stanford Microarray Database (<http://genome-www5.stanford.edu/MicroArray/SMD/>) and follow MIAME guidelines for arrays experiments [15,16]. Raw datafiles as well as the MIAME checklist are available at our laboratory webpage at <http://www.meduniwien.ac.at/nephrogene/data/DGF/>.

Microarray data analysis

The microarray dataset consisted of 41,421 cDNA features. 41,025 of those held a UniGene Cluster ID (27,442 unique genes), 396 were expressed sequence tags (ESTs) not assigned to a UniGene Cluster. Mean sector and printing plate ANOVA R^2 -values of the microarray experiments were on average 4.5×10^{-2} and 3.1×10^{-2} respectively, suggesting no dependency of results on spatial location or plate printing procedures. In a first pre-processing step a quality filter was applied on the dataset by considering only genes and ESTs with spot intensities of at least 1.5-fold over background in either channel 1 or 2 of the microarray thus leaving 32,588 cDNA features in the dataset. Only genes and ESTs with at least 80% of valid entries were considered for successive analysis steps thus further reducing the dataset to 24,624 cDNA features. The remaining missing data points were substituted applying a k-nearest-neighbor algorithm, where the number of neighbors, k, was set to ten [17]. No correction for a putative batch bias was necessary because only one array batch was used in the whole analysis for all arrays. We used the SAM methods as well as the student's t-test in order to find differentially regulated genes (DEGs) between patients experiencing DGF and the control group with primary functioning grafts [18]. The p-value threshold was set to < 0.05 with fold-change values greater than two. The number of permutations in the significance analysis of microarrays (SAM) method was set to twenty-thousand and a false discovery rate of 2.5% was selected. Differentially expressed genes were hierarchically clustered and graphically represented using the MultiExperiment Viewer developed at The Institute for Genomic Research [19]. The

cosine correlation and complete linkage were used as distance measure and linkage rule in the hierarchical cluster algorithm, respectively [19,20].

Functional data enrichment

DEGs were furthermore analyzed with respect to their molecular functions, associated biological processes, and cellular locations using gene ontology terms (GO-Terms) as provided by the Gene Ontology Consortium [21]. The SOURCE tool from the Stanford Genomics Facility was used for linking GO-Terms to the genes of interest [22]. Functional grouping of genes was based on GO-Terms, Protein ANALysis THrough Evolutionary Relationships (PANTHER) ontologies, and information derived from the protein data retrieval system iHOP [23,24].

Regulatory network analysis

All identified DEGs were mapped on a molecular dependency graph holding about 70,000 annotated human proteins [25]. Each graph node codes for a particular protein and edges between nodes encode pairwise dependencies. Dependencies were computed based on protein-protein interaction information, similarity in gene expression, conjoint regulatory patterns on the level of transcription factors and microRNAs, as well as assignment to functional ontologies. Subnetworks holding at least two DEGs were retrieved and further analyzed on a functional level.

Statistical analysis

Continuous data were analyzed by Wilcoxon rank-sum tests, categorical data by chi-square tests or Fisher's exact tests when appropriate. A p-value less than 0.05 was considered statistically significant. For all analyses SAS for Windows 9.2 (The SAS Institute, Inc., Cary, North Carolina, USA) was used.

RESULTS

Demographic data on transplant donors and recipients are provided in Table 1.

	PF group	DGF group	p-value
Number of donors	16		na
Number of donor organs	10	10	na
Donor age (years)	52.5 (45.0, 58.0)	62.5 (55.0, 72.0)	0.045
Donor sex (f/m)	4/6	7/3	0.370*
Last creatinine of donor (mg/dl)	1.00 (0.71, 1.20)	0.70 (0.60, 1.00)	0.254
Vasopressors used (n/y)	2/8	0/10	0.136
Multiorgan donors (n/y)	7/3	8/2	1.000*
Cause of death (trauma / intracranial hemorrhage / cardiac arrest / else)	1/8/1/0	0/9/0/1	0.383
Number of recipients	10	10	na
Recipient age (years)	57.3 (51.6, 62.2)	59.1 (46.3, 67.1)	0.734
Recipient sex (f/m)	3/7	3/7	1000
Transplant number (1/2)	9/1	9/1	1.000*
Cold ischemic time (hours)	9.9 (7.0, 15.0)	12.7 (10.3, 4.4)	0.308
PRA latest (%)	0.0 (0.0, 2.0)	0.0 (0.0, 2.0)	1000
Sum of HLA mismatches (0/1/2/3/4/5/6)	0/1/4/1/1/0/0	0/0/1/3/1/5/0	0.076*
Number of dialysis treatment (0/1/2/3/4)	10/0/0/0/0	3/5/0/1/1	0.003*
Immunosuppression (CNI/else)	8/2	9/1	1.000*
Induction therapy (none/antiCD25/ATG)	6/4/0	7/3/0	0.639

na ... not applicable, * Fisher's exact test

Table 1: Demographic data of transplant donors and recipients stratified by treatment assignment. Continuous data are provided as median (1st, 3rd quartile), categorical data are shown as counts.

Molecular signatures separating DGF from primary function (PF) in steroid treated donor organs

Using the SAM method sixty-three transcripts could be identified as significantly differentially regulated. Both gene lists are provided in the supplementary material (tables S1 and S2) sorted by fold-change values.

In total 147 features showed fold-change values greater than two and p-values smaller than 0.05 following a t-test. The majority of features were suppressed with only ten genes being upregulated in the DGF as compared to the PF group.

An expression profile based clustering resulted in an almost complete discrimination between DGF and PF samples as given in figure 1.

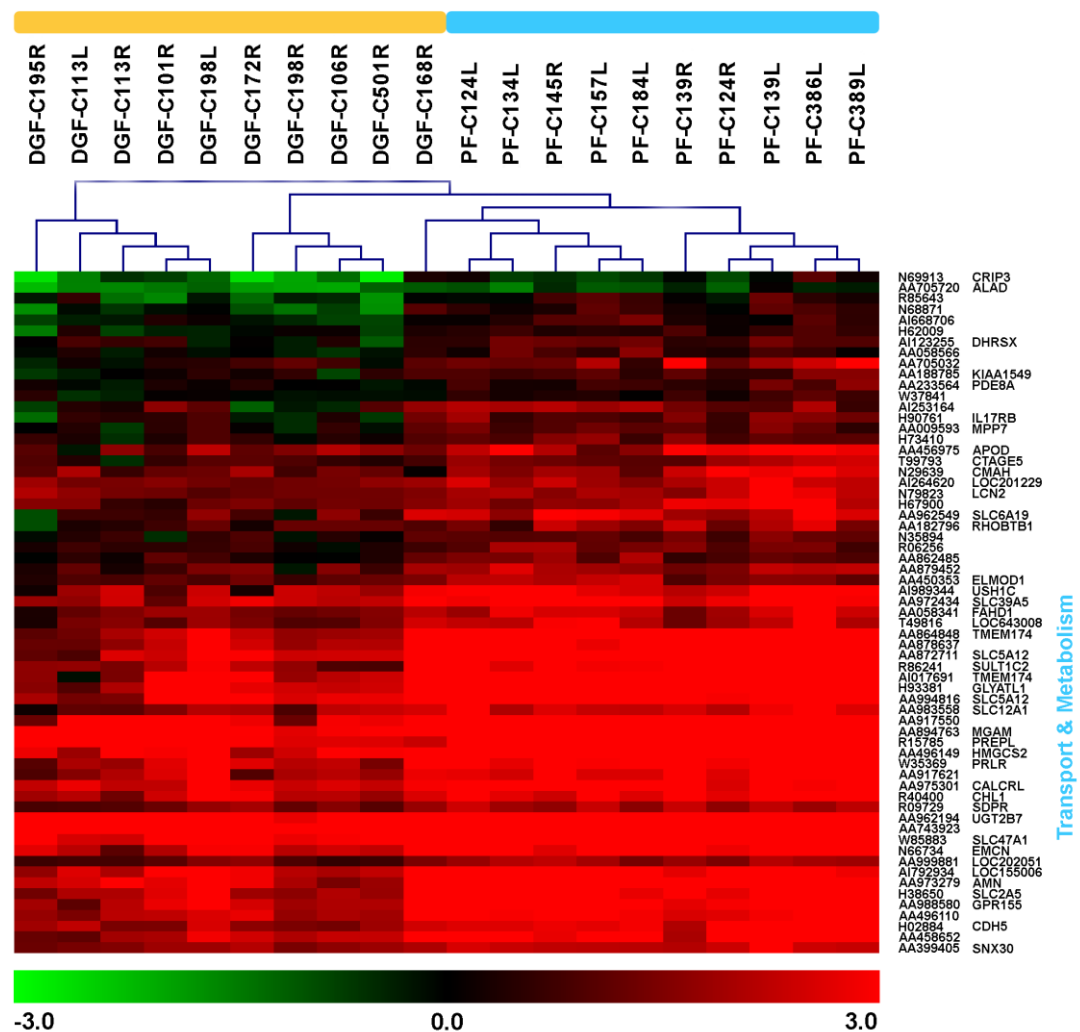


Figure 1: Dendrogram derived by unsupervised hierarchical clustering of gene expression profiles dichotomizing DGF group (orange bar) from PF (blue bar), all received steroid pretreatment. Red spots indicate upregulated transcripts, whereas green spots indicate downregulated transcripts relative to the reference RNA used. The differentially regulated genes associated with DGF could be categorized according to GO-terms mainly into Transport and Metabolism.

Functional analysis

Thirty-nine out of the 63 transcripts (SAM, 41 unique genes) and eighty-four out of the 135 downregulated transcripts (t-test, 91 unique genes) and could be mapped to PANTHER IDs. Significantly enriched or depleted biological processes with at least two members are given in table 2 (p-value < 0.05 given by a chi-square test when comparing the number of genes associated to the category with the total number of genes belonging to this particular process). Enriched processes mainly include genes involved in transport and metabolism. DGF-associated downregulated genes include many transcripts encoding solute carriers (ion, amino acid and glucose transporters) in the plasma membrane and other transporters in the cytoplasm and extracellular space. Prominent members are the organic anion transporter (SLC22A8), neutral amino acid transporter (SLC6A19), the sodium/glucose cotransporter (SLC5A12), lipocalin 2 (LCN2), and apolipoprotein D (APOD). Proteins involved in metabolism, including lipid, fatty acid, and steroid metabolism, were predominantly downregulated in DGF samples. Depleted processes are nucleoside and protein metabolism, mRNA transcription and intracellular protein traffic. Upregulated transcripts (t-test, nine unique genes) were mainly associated with blood clotting as well as immunity and defense.

Biological Process	t-test (n=84)		SAM (n=39)	
	number of genes	p-value	number of genes	p-value
DEGs down-regulated in DGF/enriched processes				
Transport	20	<0.001	8	0.001
Lipid, fatty acid and steroid metabolism	12	<0.001	5	0.006
Amino acid metabolism	7	<0.001	2	0.049
Steroid hormone metabolism	4	<0.001	2	0.002
Steroid metabolism	6	<0.001	3	0.003
Ion transport	9	<0.001	-	-
Coenzyme and prosthetic group metabolism	5	<0.001	3	0.003
Amino acid transport	3	0.001	-	-
Carbohydrate metabolism	8	0.001	-	-
Fatty acid metabolism	4	0.004	-	-
Other amino acid metabolism	2	0.005	-	-
Cation transport	6	0.005	-	-
Electron transport	4	0.010	-	-
Vitamin/cofactor transport	2	0.011	-	-

Other polysaccharide metabolism	3	0.012	-	-
Cell adhesion	6	0.017	-	-
Homeostasis	3	0.028	-	-
Extracellular transport and import	2	0.028	-	-
Anion transport	2	0.034	-	-
Sulfur metabolism	2	0.035	-	-
Proteolysis	7	0.036	-	-
Other developmental process	2	0.042	-	-
DEGs down-regulated in DGF/depleted processes				
Nucleoside, nucleotide and nucleic acid metabolism	5	0.042	-	-
Intracellular protein traffic	0	0.043	-	-
mRNA transcription	2	0.047	-	-
	t-Test (n=9)		SAM (n=0)	
DEGs up-regulated in DGF/enriched processes				
Blood circulation and gas exchange	2	<0.001	-	-
Blood clotting	2	<0.001	-	-
Immunity and defense	3	0.009	-	-

Table 2: Functional classification of DEGs using PANTHER ontologies: Enriched or depleted biological processes separating DGF and PF as derived on the level of differential gene expression by t-test and SAM. Categories are ranked by the p-value (comparison of expected number of genes and observed number of genes in each biological process) indicating the relevance of a particular process.

Interactome Analysis

We retrieved in total seven networks holding at least two of the differentially regulated genes (figure 2). Members of network cluster 1 holding 13 proteins are mainly involved in blood clotting with fibrinogen gamma (FGG), fibrinogen alpha (FGA), and the frizzled homology 8 being upregulated in patient samples experiencing DGF. Hypoxia and an older donor age might lead to the activation of fibrotic pathways which contribute to DGF. The central protein of network cluster two is the suppressor of cytokine signalling 3 (SOCS3) that shows higher expression values in the group of patients with DGF post transplant. The other network clusters contain mainly downregulated genes with members of cluster 6 being involved in steroid metabolism and members of clusters 4 and 7 being involved in lipid and fatty acid metabolism (figure 2).

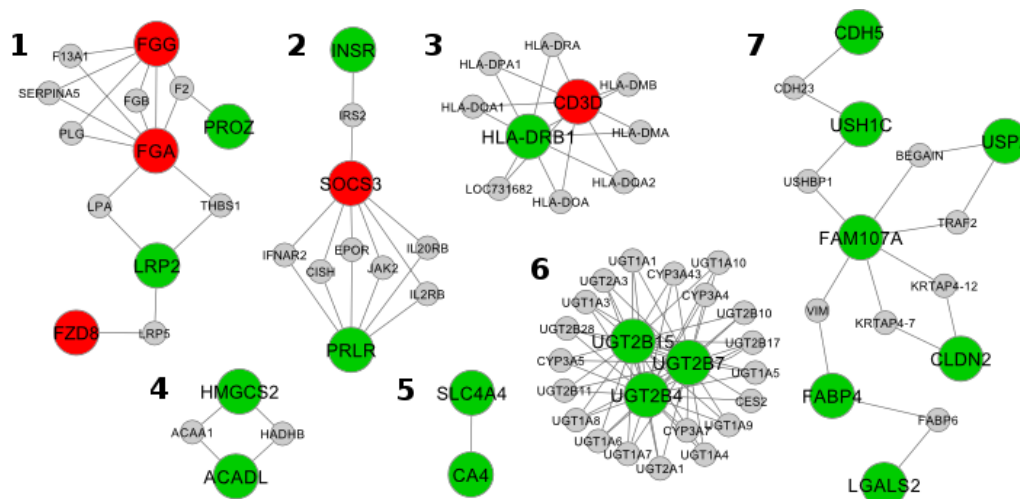


Figure 2: Seven identified networks with at least two differentially regulated genes between DGF and PF samples. Red nodes depict upregulated genes in DGF samples whereas green nodes depict downregulated genes. Differentially expressed proteins showed a high connectivity in these networks, thus indicating concerted interaction and relevance in the development of DGF.

DISCUSSION

In the present study we elucidated molecular mechanisms associated with delayed graft function after renal transplantation in zero-hour donor kidney biopsies pretreated with corticosteroids. Based on our findings poor initial function can be explained by a partial shutdown of metabolism and transport activity on a molecular level.

One possible explanation of reduced transport and metabolism is hypoxia. In the absence of oxygen severe energy depletion, i.e. less production of ATP and subsequent activation of number of critical alterations in metabolism, occurs [26]. The effects of limited oxygen supply are aggravated by the higher demand associated with the high tubular oxygen consumption necessary for solute exchange [27] and the high rate of aerobic glycolysis [28]. Hypoxia is also a profibrogenic stimulus for tubular cells, interstitial fibroblasts, and renal microvascular endothelial cells. Hypoxia can also activate fibroblasts and change the extracellular matrix metabolism of resident renal cells [29,30] and was shown to play a role in the progression of chronic kidney disease

[31]. Therefore, the use of effective preservation solutions and reduction of cold ischaemia times may improve kidney function after transplantation [32].

The downregulation of many transporters is probably caused by less oxygen supply and subsequent energy depletion. The solute carrier family 4, sodium bicarbonate cotransporter, member 4 (SLC4A4) built a small cluster with the carbonic anhydrase IV (CA4) and is involved in the regulation of bicarbonate secretion and absorption and intracellular pH suggesting tubular acidosis (figure 2). Protein-protein interactions of transporters in the molecular dependency graph are rare suggesting that these pathways are under-represented in the interactome analysis.

Lipid metabolism, fatty acid metabolism and steroid metabolism are downregulated in DGF samples and are the most enriched functional categories next to transport function (figure 2, network clusters 4, 6, 7). Although the hydroxyprostaglandin dehydrogenase 15-(NAD) (HPGD), the sulfotransferase family, cytosolic, 1C, member 2 (SULT1C2), and the three glucuronosyltransferase 2 family polypeptides UGT2B15, UGT2B4, UGT2B7 are members of the steroid metabolism they cannot be linked directly to methylprednisolone treatment. Another prominent gene, the suppressor of cytokine signaling 3 (SOCS3), belongs to a family of negative-feedback regulators of cytokine signaling. This regulator is induced by its corresponding cytokines leading to the subsequent shutdown of the respective signaling cascade [33]. SOCS3 is involved in the JAK/STAT-dependent cytokine signaling pathways and is linked to the downregulated prolactin receptor (PRLR). On the other side SOCS3 is linked over IRS2 (insulin receptor substrate 2) to the downregulated insulin receptor (INSR) (figure 2, cluster 2).

Reduced transport activity and metabolism indicating poorer quality of renal grafts was also reported by other transcriptomics studies of donor kidney biopsies developing DGF [6,34,35]. Roughly one third of reported downregulated genes by Mueller et. al. was also identified in our study strengthening the validity of obtained results. The common theme of inflammation and immune response in the context of DGF was delineated in all three studies. The suppression of inflammation with corticosteroids in our study lead to the identification of novel molecular mechanisms besides inflammation and complement activation associated with the development of DGF, namely limited transport capabilities and decreased metabolic activity of the renal organ. However, one cluster in the dependency graph with the down-regulated major histocompatibility complex, class II, DR beta 3 (HLA-DRB1) and the up-regulated CD3d molecule, delta (CD3-TCR complex) (CD3D) belongs to immunity response.

A fair number of induced genes in DGF samples could be linked to blood clotting with fibrinogen gamma and fibrinogen alpha being two prominent members. This might in part be explained by the older donors in the DGF group. Donor age is a well known risk factor of DGF but not all grafts from old donors have necessarily poor graft function. Determination of the graft quality based on demographic/clinical and molecular risk factors probably provides a much better forecast model [4]. Especially the shortage of donor organs makes an expansion of donor criteria to include older and non-heart beating donors necessary with the risk of higher rates of DGF. Therefore a better understanding of molecular mechanisms leading to DGF is of great interest and new strategies and better donor management is of great importance for the prevention of this disease.

A limitation of the present study is probably the use of cDNA arrays which cannot discriminate between different splice variants in the measurement of expression levels. Nonetheless we could identify genes mainly involved in transport and lipid, glucose metabolism associated with delayed graft function in renal transplants.

Based on these results the activation of lipid and glucose metabolism may prevent the graft from developing acute renal failure. One possible treatment strategy is the administration with peroxisome proliferator-activated receptor (PPAR) agonists. The PPARs are ligand-activated transcription factors that control lipid and glucose metabolism. Activation of PPARs negatively regulates the expression of genes induced by cerebral ischemia/reperfusion injury and was shown to prevent post-ischemic inflammation and neuronal damage in several in vitro and in vivo models [36].

Another possible strategy to revert the effects of hypoxia is the treatment with caspase inhibitors. The administration of caspase inhibitors in vivo was demonstrated to protect against cell death in animal models of ischemic acute renal failure [37]. The pancaspase inhibitor Q-VD-OPH prevents the rise in caspase activity and apoptosis [38]. Therefore PPAR-agonists and caspase inhibitors may be adopted in the donor pretreatment to prevent ischemic/reperfusion injury in the kidney. Donor pretreatment has great advantages for the recipient because improved long-term survival could thus be achieved cost-efficiently and without great effort or side effects.

In summary our analyses provide novel insight into biological processes that are associated with postischemic DGF. Based on our findings prospective trials with

targeted therapy, including PPAR-agonists or caspase inhibitors, may be designed to elucidate the causal inference of these risk markers of DGF.

Acknowledgements

This study was supported by grants from the Austrian Science Fund and the Austrian Academy of Science (FWF P-18325 to R.O.). We acknowledge the valuable contribution of the OPO coordinators.

SUPPLEMENTAL DATA

Accession No.	Gensymbol	Name	Biological Process	Fold change
AI017691	TMEM174	Transmembrane protein 174		-3.49
AA962549	SLC6A19	Solute carrier family 6 (neutral amino acid transporter), member 19	Transport, Amino acid metabolism	-3.48
AA743923		CDNA FLJ32283 fis, clone PROST2000212		-3.34
AA878637		Transcribed locus		-3.29
AA864848	TMEM174	Transmembrane protein 174		-3.20
AA962194	UGT2B7	UDP glucuronosyltransferase 2 family, polypeptide B7	Steroid hormone metabolism, Steroid metabolism, Lipid, fatty acid and steroid metabolism	-3.19
H93381	GLYATL1	Glycine-N-acyltransferase-like 1	Lipid, fatty acid and steroid metabolism	-3.10
AA872711	SLC5A12	Solute carrier family 5 (sodium/glucose cotransporter), member 12	Transport	-2.98
AI253164		Transcribed locus		-2.90
AA988580	GPR155	G protein-coupled receptor 155		-2.85
R86241	SULT1C2	Sulfotransferase family, cytosolic, 1C, member 2	Steroid hormone metabolism, Steroid metabolism, Lipid, fatty acid and steroid metabolism	-2.74

AA456975	APOD	Apolipoprotein D	Transport, Coenzyme and prosthetic group metabolism	-2.62
AA917550		Transcribed locus		-2.41
N68871		CDNA FLJ43400 fis, clone OCBBF2010281		-2.35
AA994816	SLC5A12	Solute carrier family 5 (sodium/glucose cotransporter), member 12	Transport	-2.31
AA496149	HMGCS2	3-hydroxy-3-methylglutaryl-Coenzyme A synthase 2 (mitochondrial)	Coenzyme and prosthetic group metabolism, Steroid metabolism, Lipid, fatty acid and steroid metabolism	-2.27
AA705032		Transcribed locus		-2.20
AA894763	MGAM	Maltase-glucoamylase (alpha-glucosidase)		-2.19
R15785	PREPL	Prolyl endopeptidase-like	Protein metabolism and modification	-2.18
AI792934	LOC155006	Hypothetical protein LOC155006		-2.17
W85883	SLC47A1	Solute carrier family 47, member 1		-2.15
AA973279	AMN	Amnionless homolog (mouse)	Transport, Lipid, fatty acid and steroid metabolism	-2.14
N69913	CRIP3	Cysteine-rich protein 3		-2.14
R40400	CHL1	Cell adhesion molecule with homology to L1CAM (close homolog of L1)	Determination of dorsal/ventral axis	-2.13
H02884	CDH5	Cadherin 5, type 2, VE-cadherin (vascular epithelium)		-2.11
H38650	SLC2A5	Solute carrier family 2 (facilitated glucose/fructose transporter), member 5	Transport	-2.11
AA999881	LOC202051	Hypothetical protein LOC202051		-2.07
AA972434	SLC39A5	Solute carrier family 39 (metal ion transporter), member 5	Transport	-2.05
AA058341	FAHD1	Fumarylacetoacetate hydrolase domain containing 1	Amino acid metabolism	-2.03
AA879452		CDNA clone IMAGE:5270438		-2.03
W35369	PRLR	Prolactin receptor	Lactation, mammary development	-2.01
AI989344	USH1C	Usher syndrome 1C (autosomal recessive, severe)		-2.01
N29639	CMAH	Cytidine monophosphate-N-acetylneuraminic acid hydroxylase (CMP-N-acetylneuraminate monooxygenase) pseudogene		-1.97

AA975301	CALCRL	Calcitonin receptor-like		-1.96
AA917621		Transcribed locus		-1.95
AI264620	LOC201229	Hypothetical protein LOC201229		-1.89
AA458652		Transcribed locus		-1.87
R06256		Transcribed locus		-1.85
AA496110		Transcribed locus, strongly similar to NP_115821.1 multiple EGF-like-domains 11 [Homo sapiens]		-1.84
T99793	CTAGE5	CTAGE family, member 5		-1.83
AA705720	ALAD	Aminolevulinate, delta-, dehydratase	Coenzyme and prosthetic group metabolism, Porphyrin metabolism	-1.81
AA182796	RHOBTB1	Rho-related BTB domain containing 1		-1.80
AA009593	MPP7	Membrane protein, palmitoylated 7 (MAGUK p55 subfamily member 7)	Asymmetric protein localization	-1.79
R09729	SDPR	Serum deprivation response (phosphatidylserine binding protein)	mRNA transcription termination	-1.78
AI123255	DHRX	Dehydrogenase/reductase (SDR family) X-linked		-1.78
AA450353	ELMOD1	ELMO/CED-12 domain containing 1		-1.75
AI668706		Transcribed locus		-1.75
T49816	LOC643008	PP12104		-1.73
R85643		Data not found		-1.72
H62009		Transcribed locus		-1.72
N66734	EMCN	Endomucin		-1.72
AA983558	SLC12A1	Solute carrier family 12 (sodium/potassium/chloride transporters), member 1	Transport	-1.72
AA233564	PDE8A	Phosphodiesterase 8A		-1.70
AA862485		Data not found		-1.70
N79823	LCN2	Lipocalin 2	Transport	-1.65
AA058566		Data not found		-1.65
H67900		Transcribed locus, moderately similar to XP_001372821.1 PREDICTED: similar to Choline/ethanolamine phosphotransferase 1 [Monodelphis domestica]		-1.64
H90761	IL17RB	Interleukin 17 receptor B		-1.63
AA188785	KIAA1549	KIAA1549		-1.62
H73410		Data not found		-1.62

W37841		CDNA clone IMAGE:4902949		-1.57
N35894		Data not found		-1.56
AA399405	SNX30	Sorting nexin family member 30		-1.50

Table S1. Sixty-three differentially regulated transcripts computed with the Significance Analysis of Microarrays (SAM) method sorted by fold-change values. The number of permutations in the SAM method was set to twenty-thousand and a false discovery rate of 2.5% was selected.

Accession No.	Gensymbol	Name	Biological Process	Fold change
T94626	FGG	Fibrinogen gamma chain	Blood circulation and gas exchange, Blood clotting, Immunity and defense, Cell proliferation and differentiation	4.92
AA865707	FGA	Fibrinogen alpha chain	Blood circulation and gas exchange, Blood clotting, Immunity and defense, Cell proliferation and differentiation	3.19
R14976		Data not found		3.16
AA704242	SERPINA3	Serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 3		2.84
T72915	SOCS3	Suppressor of cytokine signaling 3	JAK-STAT cascade, Inhibition of apoptosis	2.72
AA457138	FZD8	Frizzled homolog 8 (Drosophila)		2.54
AI003775	LOC387763	Hypothetical LOC387763		2.48
H53340	MT1G	Metallothionein 1G		2.44
AW029498	SERPINA3	Serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 3		2.42
AI922872	SOCS3	Suppressor of cytokine signaling 3		2.42
AA055946	CD3D	CD3d molecule, delta (CD3-TCR complex)	Immunity and defense	2.12
AA678021	SNRPE	Small nuclear ribonucleoprotein polypeptide E		2.01

AA704995	GLYAT	Glycine-N-acyltransferase	Lipid, fatty acid and steroid metabolism, Fatty acid metabolism	-4.71
AI253049	TINAG	Tubulointerstitial nephritis antigen	Cell adhesion	-4.01
AA932134		CDNA FLJ32283 fis, clone PROST2000212		-3.90
R97050		CDNA clone IMAGE:4610527		-3.78
AA994816	SLC5A12	Solute carrier family 5 (sodium/glucose cotransporter), member 12	Transport	-3.68
AA877253	RNF186	Ring finger protein 186	Proteolysis	-3.63
AA885603		Transcribed locus		-3.52
AI017691	TMEM174	Transmembrane protein 174		-3.50
AA962549	SLC6A19	Solute carrier family 6 (neutral amino acid transporter), member 19	Transport, Amino acid metabolism, Amino acid transport	-3.48
W85851	ACSM2B	Acyl-CoA synthetase medium-chain family member 2B		-3.38
AA743923		CDNA FLJ32283 fis, clone PROST2000212		-3.34
AA878637		Transcribed locus		-3.29
AA864848	TMEM174	Transmembrane protein 174		-3.21
AA919149	HAO2	Hydroxyacid oxidase 2 (long chain)	Carbohydrate metabolism	-3.20
AA962194	UGT2B7	UDP glucuronosyltransferase 2 family, polypeptide B7	Lipid, fatty acid and steroid metabolism, Steroid hormone metabolism, Steroid metabolism, Carbohydrate metabolism, Other polysaccharide metabolism	-3.19
N74025	DIO1	Deiodinase, iodothyronine, type I		-3.17
AI017796	SLC5A12	Solute carrier family 5 (sodium/glucose cotransporter), member 12	Transport	-3.15
H93381	GLYATL1	Glycine-N-acyltransferase-like 1	Lipid, fatty acid and steroid metabolism, Fatty acid metabolism	-3.09
H88329	CALB1	Calbindin 1, 28kDa	Homeostasis	-3.04
AA872711	SLC5A12	Solute carrier family 5 (sodium/glucose cotransporter), member 12	Transport	-2.98
AI017796	SLC5A12	Solute carrier family 5 (sodium/glucose cotransporter), member 12	Transport	-2.96
AI335086	ANGPTL3	Angiopoietin-like 3		-2.93

AI245843		Transcribed locus, strongly similar to NP_001011880.1 hypothetical protein LOC497190 [Homo sapiens]		-2.91
AI253164		Transcribed locus		-2.90
AI264674	SLC16A12	Solute carrier family 16, member 12 (monocarboxylic acid transporter 12)	Transport, Ion transport, Cation transport	-2.88
AA988580	GPR155	G protein-coupled receptor 155		-2.85
R08178	LOC100129488	Hypothetical protein LOC100129488		-2.85
AA864183	RHCG	Rh family, C glycoprotein	Transport	-2.82
AA928710	SLC6A19	Solute carrier family 6 (neutral amino acid transporter), member 19	Transport, Amino acid metabolism, Amino acid transport	-2.74
R86241	SULT1C2	Sulfotransferase family, cytosolic, 1C, member 2	Lipid, fatty acid and steroid metabolism, Steroid hormone metabolism, Steroid metabolism, Sulfur metabolism	-2.74
AA456001	NOX4	NADPH oxidase 4	Electron transport	-2.70
N36136	EMCN	Endomucin	Cell adhesion	-2.69
AA416585	ACE2	Angiotensin I converting enzyme (peptidyl-dipeptidase A) 2	Proteolysis	-2.68
AI241028		Data not found		-2.68
AA994857	ZNF552	Zinc finger protein 552	Nucleoside, nucleotide and nucleic acid metabolism, mRNA transcription	-2.67
AA514359	RNF186	Ring finger protein 186	Proteolysis	-2.66
AA456975	APOD	Apolipoprotein D	Transport, Coenzyme and prosthetic group metabolism, Vitamin/cofactor transport	-2.62
AI301528	HNF4A	Hepatocyte nuclear factor 4, alpha		-2.60
T70353	ACMSD	Aminocarboxymuconate semialdehyde decarboxylase		-2.58
N53031	UGT2B4	UDP glucuronosyltransferase 2 family, polypeptide B4	Lipid, fatty acid and steroid metabolism, Steroid hormone metabolism, Steroid metabolism, Carbohydrate metabolism, Other polysaccharide metabolism	-2.57
W81603		Data not found		-2.56
AA902897		Transcribed locus		-2.53
R16259		Data not found		-2.53
H44449	LRP2	Low density lipoprotein-related protein 2		-2.51
AA878939		Transcribed locus		-2.51

R63647	PRLR	Prolactin receptor	Lactation, mammary development	-2.49
H18608	SLC22A8	Solute carrier family 22 (organic anion transporter), member 8	Transport, Ion transport, Extracellular transport and import, Anion transport	-2.45
AI245812	KCNJ15	Potassium inwardly-rectifying channel, subfamily J, member 15	Transport, Ion transport, Cation transport	-2.42
AA918008	SLC28A1	Solute carrier family 28 (sodium-coupled nucleoside transporter), member 1	Transport, Ion transport, Cation transport, Nucleoside, nucleotide and nucleic acid metabolism	-2.42
AA932134		CDNA FLJ32283 fis, clone PROST2000212		-2.42
AA932135		Transcribed locus		-2.41
AA917550		Transcribed locus		-2.41
AI015991	CLDN2	Claudin 2		-2.40
AA746229	UGT2B7	UDP glucuronosyltransferase 2 family, polypeptide B7	Lipid, fatty acid and steroid metabolism, Steroid hormone metabolism, Steroid metabolism, Carbohydrate metabolism, Other polysaccharide metabolism	-2.40
T50951	UGT2B15	UDP glucuronosyltransferase 2 family, polypeptide B15	Lipid, fatty acid and steroid metabolism, Steroid hormone metabolism, Steroid metabolism, Carbohydrate metabolism, Other polysaccharide metabolism	-2.39
N53031	UGT2B4	UDP glucuronosyltransferase 2 family, polypeptide B4	Lipid, fatty acid and steroid metabolism, Steroid hormone metabolism, Steroid metabolism, Carbohydrate metabolism, Other polysaccharide metabolism	-2.38
AI222515	BBOX1	Butyrobetaine (gamma), 2-oxoglutarate dioxygenase (gamma-butyrobetaine hydroxylase) 1	Coenzyme and prosthetic group metabolism	-2.37
AI264674	SLC16A12	Solute carrier family 16, member 12 (monocarboxylic acid transporter 12)	Transport, Ion transport, Cation transport	-2.35
N68871		CDNA FLJ43400 fis, clone OCBBF2010281		-2.35
R98936	MME	Membrane metallo-endopeptidase	Proteolysis	-2.34
AI261833	SLC7A9	Solute carrier family 7 (cationic amino acid transporter, y+ system), member 9	Transport, Amino acid metabolism, Amino acid transport	-2.34
AA878391	GPC5	Glypican 5	Cell adhesion	-2.31

AA994816	SLC5A12	Solute carrier family 5 (sodium/glucose cotransporter), member 12	Transport	-2.31
R43597		Data not found		-2.31
AA918729		Transcribed locus		-2.30
R08912		Data not found		-2.30
AA703222		CDNA FLJ12088 fis, clone HEMBB1002545		-2.30
AA676742	DMGDH	Dimethylglycine dehydrogenase	Electron transport	-2.27
AA496149	HMGCS2	3-hydroxy-3-methylglutaryl-Coenzyme A synthase 2 (mitochondrial)	Lipid, fatty acid and steroid metabolism, Steroid metabolism, Coenzyme and prosthetic group metabolism	-2.27
W56753	KIAA1276	KIAA1276 protein		-2.27
AA947621	ATP6V1G3	ATPase, H ⁺ transporting, lysosomal 13kDa, V1 subunit G3	Transport, Ion transport, Cation transport, Nucleoside, nucleotide and nucleic acid metabolism	-2.26
AA858019	SLC13A1	Solute carrier family 13 (sodium/sulfate symporters), member 1	Transport, Ion transport, Cation transport	-2.26
AA862436	FAM151A	Family with sequence similarity 151, member A		-2.25
R10885	ACY3	Aspartoacylase (aminocyclase) 3	Amino acid metabolism, Other amino acid metabolism	-2.24
AA287032	TBC1D8B	TBC1 domain family, member 8B (with GRAM domain)		-2.24
R98070		Data not found		-2.23
R40176	CXCL14	Chemokine (C-X-C motif) ligand 14		-2.22
AA971563	SGSM3	Small G protein signaling modulator 3		-2.22
AA026754	SNTA1	Syntrophin, alpha 1 (dystrophin-associated protein A1, 59kDa, acidic component)		-2.22
AI253036		Transcribed locus		-2.20
AA705032		Transcribed locus		-2.20
AA894763	MGAM	Maltase-glucoamylase (alpha-glucosidase)		-2.19
AI344372	SLC26A7	Solute carrier family 26, member 7	Transport, Ion transport, Extracellular transport and import, Anion transport, Sulfur metabolism	-2.19
T47312	INSR	Insulin receptor	Carbohydrate metabolism, Regulation of carbohydrate metabolism, Other developmental process	-2.18

AI792934	LOC155006	Hypothetical protein LOC155006		-2.18
AA705112	MOCS1	Molybdenum cofactor synthesis 1	Coenzyme and prosthetic group metabolism, Pterin metabolism	-2.18
R15785	PREPL	Prolyl endopeptidase-like	Proteolysis	-2.18
AI015652	SLC13A1	Solute carrier family 13 (sodium/sulfate symporters), member 1	Transport, Ion transport, Cation transport	-2.18
W85883	SLC47A1	Solute carrier family 47, member 1		-2.15
AA971425	USP2	Ubiquitin specific peptidase 2	Proteolysis	-2.15
AA973279	AMN	Amnionless homolog (mouse)	Transport, Lipid, fatty acid and steroid metabolism	-2.14
AA677185	ANK3	Ankyrin 3, node of Ranvier (ankyrin G)		-2.14
AI733138	BHMT2	Betaine-homocysteine methyltransferase 2	Amino acid metabolism	-2.14
AA886349		Data not found		-2.14
R66006	ACADL	Acyl-Coenzyme A dehydrogenase, long chain	Lipid, fatty acid and steroid metabolism, Fatty acid metabolism, Electron transport	-2.13
R40400	CHL1	Cell adhesion molecule with homology to L1CAM (close homolog of L1)	Cell adhesion	-2.13
N69913	CRIP3	Cysteine-rich protein 3		-2.13
N92901	FABP4	Fatty acid binding protein 4, adipocyte	Transport, Lipid, fatty acid and steroid metabolism, Coenzyme and prosthetic group metabolism, Vitamin/cofactor transport	-2.13
H50623	HLA-DRB1	Major histocompatibility complex, class II, DR beta 3		-2.13
H27752	AQP7	Aquaporin 7	Transport, Homeostasis	-2.12
AA256291		Transcribed locus		-2.12
AI263210		Transcribed locus		-2.12
H02884	CDH5	Cadherin 5, type 2, VE- cadherin (vascular epithelium)	Cell adhesion	-2.11
H38650	SLC2A5	Solute carrier family 2 (facilitated glucose/fructose transporter), member 5	Transport, Carbohydrate metabolism	-2.11
AA865572		Transcribed locus		-2.11
AA111975	CMBL	Carboxymethylenebutenolidase homolog (Pseudomonas)	Carbohydrate metabolism	-2.10
AA775223	HPGD	Hydroxyprostaglandin dehydrogenase 15-(NAD)	Lipid, fatty acid and steroid metabolism, Steroid metabolism	-2.10
AA485893	RNASE1	Ribonuclease, RNase A family, 1 (pancreatic)	Nucleoside, nucleotide and nucleic acid metabolism	-2.10

AI815076	SLC7A7	Solute carrier family 7 (cationic amino acid transporter, y+ system), member 7	Transport, Amino acid metabolism, Amino acid transport	-2.10
W72294	CXCL14	Chemokine (C-X-C motif) ligand 14		-2.09
H78003	IYD	Iodotyrosine deiodinase	Electron transport	-2.09
H18456	LOC644662	Similar to hCG2042541		-2.09
AA682293	PAH	Phenylalanine hydroxylase	Amino acid metabolism, Other amino acid metabolism	-2.09
R07484		Data not found		-2.09
AA452278	SLC4A4	Solute carrier family 4, sodium bicarbonate cotransporter, member 4	Transport, Ion transport, Cation transport, Homeostasis	-2.08
AA677050	AFM	Afamin	Transport	-2.07
AA999881	LOC202051	Hypothetical protein LOC202051		-2.07
AI279830	PPP1R16B	Protein phosphatase 1, regulatory (inhibitor) subunit 16B		-2.07
AA855158	CA4	Carbonic anhydrase IV		-2.06
AI383171	LDB3	LIM domain binding 3	Nucleoside, nucleotide and nucleic acid metabolism, mRNA transcription, Other developmental process	-2.06
AA452826	PCP4	Purkinje cell protein 4		-2.06
AA972434	SLC39A5	Solute carrier family 39 (metal ion transporter), member 5	Transport, Ion transport	-2.05
AI300876	FAM150B	Family with sequence similarity 150, member B		-2.04
AA058341	FAHD1	Fumarylacetoacetate hydrolase domain containing 1	Amino acid metabolism	-2.03
AA932696	FAM107A	Family with sequence similarity 107, member A		-2.03
AA872397	LGALS2	Lectin, galactoside-binding, soluble, 2	Cell adhesion	-2.03
AI000188	UGT2B7	UDP glucuronosyltransferase 2 family, polypeptide B7	Lipid, fatty acid and steroid metabolism, Steroid hormone metabolism, Steroid metabolism, Carbohydrate metabolism, Other polysaccharide metabolism	-2.03
W35369	PRLR	Prolactin receptor	Lactation, mammary development	-2.02
AA680349	PROZ	Protein Z, vitamin K-dependent plasma glycoprotein	Proteolysis	-2.02
AA879452		CDNA clone IMAGE:5270438		-2.02

H99932	CRYL1	Crystallin, lambda 1	Lipid, fatty acid and steroid metabolism, Carbohydrate metabolism, Fatty acid metabolism	-2.01
H02824	LYVE1	Lymphatic vessel endothelial hyaluronan receptor 1		-2.01
AA579186	TMPRSS2	Transmembrane protease, serine 2	Proteolysis	-2.01
R68997	PRLR	Prolactin receptor	Lactation, mammary development	-2.00
AI989344	USH1C	Usher syndrome 1C (autosomal recessive, severe)		-2.00

Table S2. 147 differentially regulated transcripts computed with the student's t-Test sorted by fold-change values. The p-value threshold was set to < 0.05 with fold-change values greater than two.

REFERENCES

1. Perico N, Cattaneo D, Sayegh MH, Remuzzi G. Delayed graft function in kidney transplantation. *Lancet* **2004**; 364:1814-1827.
2. Ojo AO, Wolfe RA, Held PJ, Port FK, Schmouder RL. Delayed graft function: risk factors and implications for renal allograft survival. *Transplantation* **1997**; 63:968-974.
3. Schwarz C, Oberbauer R. The influence of organ donor factors on early allograft function. *Curr Opin Urol* **2003**; 13:99-104.
4. Koppelstaetter C, Schratzberger G, Perco P, et al. Markers of cellular senescence in zero hour biopsies predict outcome in renal transplantation. *Aging Cell* **2008**; 7:491-497.
5. Perco P, Kainz A, Wilflingseder J, Soleiman A, Mayer B, Oberbauer R. Histogenomics: association of gene expression patterns with histological parameters in kidney biopsies. *Transplantation* **2009**; 87:290-295.
6. Hauser P, Schwarz C, Mitterbauer C, et al. Genome-wide gene-expression patterns of donor kidney biopsies distinguish primary allograft function. *Lab. Invest* **2004**; 84:353-361.
7. Bruinsma GJ, Nederhoff MG, Geertman HJ, et al. Acute increase of myocardial workload, hemodynamic instability, and myocardial histological changes induced by brain death in the cat. *J. Surg. Res* **1997**; 68:7-15.
8. Novitzky D. Detrimental effects of brain death on the potential organ donor. *Transplant. Proc* **1997**; 29:3770-3772.
9. Power BM, Van Heerden PV. The physiological changes associated with brain death--current concepts and implications for treatment of the brain dead organ donor. *Anaesth Intensive Care* **1995**; 23:26-36.
10. Legrand M, Mik EG, Johannes T, Payen D, Ince C. Renal hypoxia and dysoxia after reperfusion of the ischemic kidney. *Mol. Med* **2008**; 14:502-516.
11. Kainz A, Mitterbauer C, Hauser P, et al. Alterations in gene expression in cadaveric vs. live donor kidneys suggest impaired tubular counterbalance of oxidative stress at implantation. *Am. J. Transplant* **2004**; 4:1595-1604.
12. Kainz A, Wilflingseder J, Mitterbauer C, et al. Steroid pretreatment of organ donors to prevent postischemic renal allograft failure: a randomized, controlled trial. *Ann. Intern. Med* **2010**; 153:222-230.
13. Groenewoud AF, Thorogood J. Current status of the Eurotransplant randomized multicenter study comparing kidney graft preservation with histidine-tryptophan-ketoglutarate, University of Wisconsin, and Euro-Collins solutions. The HTK Study Group. *Transplant. Proc* **1993**; 25:1582-1585.
14. Perou CM, Sørli T, Eisen MB, et al. Molecular portraits of human breast tumours. *Nature* **2000**; 406:747-752.

15. Brazma A, Hingamp P, Quackenbush J, et al. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet* **2001**; 29:365-371.
16. Gollub J, Ball CA, Binkley G, et al. The Stanford Microarray Database: data access and quality assessment tools. *Nucleic Acids Res* **2003**; 31:94-96.
17. Troyanskaya O, Cantor M, Sherlock G, et al. Missing value estimation methods for DNA microarrays. *Bioinformatics* **2001**; 17:520-525.
18. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. U.S.A* **2001**; 98:5116-5121.
19. Saeed AI, Sharov V, White J, et al. TM4: a free, open-source system for microarray data management and analysis. *BioTechniques* **2003**; 34:374-378.
20. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U.S.A* **1998**; 95:14863-14868.
21. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet* **2000**; 25:25-29.
22. Diehn M, Sherlock G, Binkley G, et al. SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data. *Nucleic Acids Res* **2003**; 31:219-223.
23. Mi H, Lazareva-Ulitsky B, Loo R, et al. The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Res* **2005**; 33:D284-288.
24. Hoffmann R, Valencia A. A gene network for navigating the literature. *Nat. Genet* **2004**; 36:664.
25. Bernthaler A, Mühlberger I, Fehete R, Perco P, Lukas A, Mayer B. A dependency graph approach for the analysis of differential gene expression profiles. *Mol Biosyst* **2009**; 5:1720-1731.
26. Devarajan P. Update on mechanisms of ischemic acute kidney injury. *J. Am. Soc. Nephrol* **2006**; 17:1503-1520.
27. Rosen S, Epstein FH, Brezis M. Determinants of intrarenal oxygenation: factors in acute renal failure. *Ren Fail* **1992**; 14:321-325.
28. Cohen JJ. Is the function of the renal papilla coupled exclusively to an anaerobic pattern of metabolism? *Am. J. Physiol* **1979**; 236:F423-433.
29. Norman JT, Clark IM, Garcia PL. Hypoxia promotes fibrogenesis in human renal fibroblasts. *Kidney Int* **2000**; 58:2351-2366.
30. Norman JT, Orphanides C, Garcia P, Fine LG. Hypoxia-induced changes in extracellular matrix metabolism in renal cells. *Exp. Nephrol* **1999**; 7:463-469.
31. Rudnicki M, Perco P, Enrich J, 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.

32. Moers C, Smits JM, Maathuis M-HJ, et al. Machine perfusion or cold storage in deceased-donor kidney transplantation. *N. Engl. J. Med* **2009**; 360:7-19.
33. Heeg K, Dalpke A. TLR-induced negative regulatory circuits: role of suppressor of cytokine signaling (SOCS) proteins in innate immunity. *Vaccine* **2003**; 21 Suppl 2:S61-67.
34. Mas VR, Archer KJ, Yanek K, et al. Gene expression patterns in deceased donor kidneys developing delayed graft function after kidney transplantation. *Transplantation* **2008**; 85:626-635.
35. Mueller TF, Reeve J, Jhangri GS, et al. The transcriptome of the implant biopsy identifies donor kidneys at increased risk of delayed graft function. *Am. J. Transplant* **2008**; 8:78-85.
36. Collino M, Patel NSA, Thiernemann C. PPARs as new therapeutic targets for the treatment of cerebral ischemia/reperfusion injury. *Ther Adv Cardiovasc Dis* **2008**; 2:179-197.
37. Melnikov VY, Faubel S, Siegmund B, Lucia MS, Ljubanovic D, Edelstein CL. Neutrophil-independent mechanisms of caspase-1- and IL-18-mediated ischemic acute tubular necrosis in mice. *J. Clin. Invest* **2002**; 110:1083-1091.
38. Jani A, Ljubanovic D, Faubel S, Kim J, Mischak R, Edelstein CL. Caspase inhibition prevents the increase in caspase-3, -2, -8 and -9 activity and apoptosis in the cold ischemic mouse kidney. *Am. J. Transplant* **2004**; 4:1246-1254.

2.3.1 The Thesis Author's Contribution

The thesis author performed parts of the functional enrichment analysis of differentially expressed genes in kidney grafts after steroid pretreatment and contributed to the selection of relevant functional categories. The discussion of results from the network analysis was jointly conducted by all of the authors.

In detail, the following contributions are due to the thesis author's efforts:

- Functional classification of differentially expressed genes with respect to biological processes using the PANTHER classification tool
- Contributions to the interpretation of results derived from the bioinformatics analyses, namely enriched biological processes and protein networks
- Provision of bioinformatics-specific methods and results sections to the manuscript draft

2.4 Linking transcriptomic and proteomic data on the level of protein interaction networks. Electrophoresis. 2010

Linking transcriptomics and proteomics data on the level of protein interaction networks

Paul Perco¹, Irmgard Mühlberger¹, Gert Mayer², Rainer Oberbauer³, Arno

Lukas¹, Bernd Mayer^{1,4 *}

¹ emergentec biodevelopment GmbH, Rathausstrasse 5/3, 1010 Vienna, Austria

² Medical University of Innsbruck, Department of Internal Medicine IV, Anichstrasse 35, 6020 Innsbruck, Austria

³ Medical University of Vienna, Department of Internal Medicine III, Waehringer Guertel 18-20, 1090 Vienna, Austria

⁴ Institute for Theoretical Chemistry, University of Vienna, Währingerstrasse 17, 1090 Vienna, Austria

* Corresponding author:

Dr. Bernd Mayer
emergentec biodevelopment GmbH
Gersthofen Strasse 29-31
1180 Vienna, Austria
phone: +43-1-4034966
fax: +43-1-4034966-19
e-mail: bernd.mayer@univie.ac.at; bernd.mayer@emergentec.com

Published in: Electrophoresis. 2010 Jun;31(11):1780-9.

LIST OF ABBREVIATIONS

2D-PAGE – Two Dimensional Poly Acrylamid Gel Electrophoresis

CE – Capillary Electrophoresis

CKD – chronic kidney disease

DAVID – Database for Annotation, Visualization, and Integrated Discovery

ECM – extracellular matrix

HPLC – High Performance Liquid Chromatography

HUPDB – Human Urinary Proteome Database

KEGG – Kyoto Encyclopedia of Genes and Genomes

MAPPER – Multi-genome Analysis of Positions and Patterns of Elements of Regulation

PANTHER – Protein Analysis THrough Evolutionary Relationships

PRIDE – Proteomics IDentification database

ABSTRACT

Integration and joint analysis of omics profiles derived on the genome, transcriptome, proteome and metabolome level is a natural next step in realizing a Systems Biology view of cellular processes. However, merging e.g. mRNA concentration and protein abundance profiles is not straight forward, as a direct overlap of differentially regulated/abundant features resulting from transcriptomics and proteomics is for various reasons limited. We present procedures for integrating omics profiles at the level of protein interaction networks, exemplified by using transcriptomics and proteomics data sets characterizing chronic kidney disease.

On the level of direct feature overlap only a limited number of genes and proteins were found to be significantly affected following a separate transcript and protein profile analysis, including a collagen subtype and uromodulin, both being described in the context of renal failure. On the level of protein pathway and process categories this minor overlap increases substantially, identifying cell structure, cell adhesion, as well as immunity and defense mechanisms as jointly populated with features individually identified as relevant in transcriptomics and proteomics experiments.

Mapping diverse data sources characterizing a given phenotype under analysis on directed but also undirected protein interaction networks serves in joint functional interpretation of omics data sets.

INTRODUCTION

High-throughput transcriptomics and proteomics experiments have paved the way in molecular biology research to study thousands of cellular components in parallel [1-3]. Gene Chips from e.g. Affymetrix cover roughly 29,000 human open reading frames. Gene expression profiles for over 340,000 samples are currently stored in the Gene Expression Omnibus, a microarray repository hosted by the National Center for Biotechnology Information [4]. In proteomics comparable steps have been made towards large scale analysis. Here, reduction of sample complexity by separation techniques has been elaborated, mainly including HPLC, CE, and 2D-PAGE. Subsequently mass spectrometric techniques, together with computational analysis have been applied for protein identification and quantitation. Proteomics repositories have been established as e.g. PRIDE (www.ebi.ac.uk/pride), and both, proteomics as well as transcriptomics data repositories follow data standards for enabling standardized retrieval and analysis.

However, most analysis performed is 'within a domain', i.e. transcriptomics and proteomics analysis follows established workflows aimed at deriving abundance profiles where the features are ranked by statistical criteria as the significance of a fold change in a group comparison. Tackling a given hypothesis by both, transcriptomics and proteomics in parallel (ideally using the same sample source), is unfortunately done less frequent. However, utilizing resources as the Gene Expression Omnibus and PRIDE allows extracting both data levels for a number of cellular conditions, in principal enabling joint analysis of both profiles characterizing a specific phenotype. Certainly, intrinsic heterogeneity has to be respected by such an approach including deviating phenotype definition regarding cases and controls, and intrinsic experimental biases.

The general question regarding the correlation between mRNA abundance and the concentration on the protein level has been heavily discussed in the literature. One of the first studies to compare mRNA levels and protein concentrations on a global level was conducted by Gygi and colleagues in 1999 using *Saccharomyces cerevisiae* as model organism [5]. By comparing serial analysis of gene expression mRNA counts with levels of protein abundance as derived by 2D-PAGE the authors concluded that a simple deduction of protein concentrations from mRNA transcript analysis is insufficient. As major reasons for the poor correlation regulatory mechanisms during the gene expression process, post-translational modifications and protein degradation, as well as mechanisms independent of the gene expression process were identified.

Koji and colleagues found a positive correlation but concluded that mRNA abundance is not a predictor of protein abundance, as a number of high abundant transcripts were not detected on the protein level [6]. More specific numbers are provided by Lu and colleagues, reporting that 73% of the variance in yeast protein abundance is explained by mRNA concentration [7]. In a recent study by Shankavaram and colleagues utilizing a large NCI-60 cancer cell panel, around 65% of the genes in the dataset showed statistically significant transcript-protein correlation [8]. Rogers and colleagues developed a probabilistic clustering model and analyzed time-series of transcriptomics and proteomics data from a human breast epithelial cell line [9]. They found that high correlations are mainly found in specific molecular machines as cell adhesion and protein folding complexes.

Reasons for a poor correlation between mRNA and protein abundance are manifold, including regulatory mechanisms in the course of gene expression (e.g. miRNA interactions), as well as post-translational modifications altering protein half-life. On top pathophysiological mechanisms can result in high amount of protein in specific tissues although the protein synthesis rate in this specific tissue is not altered [10]. A prototypical example is the prevalence of protein in urine in chronic kidney disease caused by leakage in the tubular barrier function of the kidney.

Furthermore, depending on the detection method used, technical bias and noise in high-throughput experiments can have significant influences, as outlined by Greenbaum and colleagues who reported a correlation coefficient of 0.66 when analyzing merged proteomics and transcriptomics datasets [11]. The same group reported higher correlation coefficients of up to 0.8 for specific subsets of genes based on subcellular location or functional grouping instead of analyzing on the level of individual genes [11,12].

In summary, next to the mRNA abundance level various other factors influence effective protein concentration. With respect to the above mentioned reasons, a simple correlation between quantities of individual mRNAs and proteins is insufficient to explain the causative dependencies of these two entities. However, features identified on either transcript or protein level may at least share the same functional context. From this, the analysis of transcriptomics and proteomics data on the level of protein interaction networks (PIN) may be a way for identifying the link between such profiles.

PINs are either directed graphs as given in KEGG [13], or undirected graphs as e.g. provided in OPHID [14]. Mapping omics profiles on such graphs may identify up- or downstream links between a change in transcript abundance and consequential, non-direct change in abundance of a protein.

However, information on links between proteins is far from complete. KEGG e.g. presently represents 4756 unique genes. For overcoming this limitation we have recently developed omicsNET aimed at linking gene/protein lists resulting from omics experiments on the level of a complete protein dependency network [15]. This protein dependency graph holds pair-wise dependencies for all presently annotated human protein-coding genes.

In the current study we compare and analyze transcriptomics and proteomics profiles reported in the context of chronic kidney disease (CKD).

Chronic kidney disease is a major clinical issue with around 10% of the population in western industrialized countries being affected according to recent reports [16]. CKD is classified into stages based on the level of the glomerular filtration rate (GFR), which normally is approximately 120 - 130 ml/min/1.73 m² with considerable variation between and even within individuals. Below 60 ml/min/1.73 m² the rate of complications based on filtration inefficiency increases, and the risk of cardiovascular events is elevated even at earlier stages. The most severe form of CKD is end stage renal disease, resulting in dialysis or transplantation as only therapy options. Transcriptomics as well as proteomics methodologies have significantly contributed towards unraveling molecular mechanisms leading to CKD [17-19], and linking available omics profiles promises further understanding of this disease.

MATERIALS AND METHODS

Data sets

We used three publicly available microarray studies on chronic kidney disease for identifying deregulated features on the mRNA level, all using kidney tissue biopsy

material. Two studies focused on differences in mRNA expression in diabetic nephropathy using Affymetrix Gene Chips. In a first study Schmid and colleagues compared mRNA levels in the tubulointerstitial compartment of thirteen diseased patients and seven healthy control subjects. The list of differentially expressed genes is provided as supplementary material with the publication [20]. The second dataset can be accessed through the Gene Expression Omnibus database (GSE1009) and was published by Baelde and colleagues. It holds transcripts differentially expressed between cells of glomeruli from diseased and morphologically normal kidneys [21]. The third study by performed by Rudnicki and colleagues on cDNA arrays identified transcripts differentially expressed between renal proximal tubular epithelial cells from biopsies of patients with nondiabetic nephropathies (IgA-nephritis, focal segmental glomerulosclerosis, and minimal-change disease) and healthy controls, respective relevant features are provided in [17].

The proteomics dataset was extracted from the Human Urinary Proteome Database v2.0 (HUPDB v2.0) available at http://mosaiques-diagnostics.de/diapatpcms/mosaiquescms/front_content.php?idcat=257, database status as of September 2009. This database holds information on protein abundance of currently 3687 human urine samples as detected by capillary electrophoresis – mass spectrometry (CE-MS) [22]. The samples were derived from patients covering a wide spectrum of different pathophysiological conditions, among them renal disorders, as well as from healthy controls. For our analysis we extracted a total of 192 samples associated with diabetic nephropathy (n=67), IgA nephropathy (n=44), membranous glomerulonephritis (n=31), focal segmental glomerulosclerosis (n=25), and minimal change disease (n=25). Experimental identification of these features was following high resolution capillary electrophoresis coupled with mass spectrometry. Certainly, chronic kidney disease itself shows various etiologies, but it is speculated that independent of the primary cause for kidney damage unified molecular processes may be seen with altered tubules. Still, numerous features identified for the 192 samples included appear sporadic (patient specific), and we decided to only select features present in at least 30% of diseased samples as being relevant. Further single proteomics studies for the given phenotype are available in the literature; however, we decided to only include samples retrieved from the HUPD as single source for not further increasing heterogeneity of data sets based on different experimental procedures used.

Analysis procedures

Differentially regulated transcripts and proteins were mapped to their respective NCBI Gene Symbols for aligning the transcriptomics and the proteomics name spaces. Since HUPD uses Swissprot names as identifier, the mapping procedure from proteins to Gene Symbols was performed using the annotation tool provided by Swissprot [23].

In a first analysis step those features present in both, the transcriptomics and proteomics list were identified. In successive analyses the overlap of lists was interpreted on the level of functional annotation, molecular pathways and protein dependency networks.

Functional annotation

Enriched biological processes based on both the transcriptomics and the proteomics list were identified using the PANTHER (Protein Analysis THrough Evolutionary Relationships) Classification System [24]. In the PANTHER ontology proteins are classified into families and subfamilies of shared function, which are further assigned to specific ontology terms in the two main categories 'biological process' and 'molecular function'. A chi-square test was used in order to identify significantly enriched or depleted biological categories when using the fully annotated set of human genes as reference dataset. Biological processes showing p-values below 0.05 were considered as statistically significant.

Pathway analysis

Pathway analysis was performed using the DAVID (Database for Annotation, Visualization, and Integrated Discovery) tool which provides gene-specific functional data mining tools and methods for functional category enrichment analysis [25][26]. The enrichment of transcripts and proteins in Kyoto Encyclopedia of Gene and Genomes (KEGG) pathways was calculated using a modified Fisher exact test. Pathways with p-values below 0.05 were considered as statistically significant.

omicsNET protein dependency network

The protein dependency analysis framework omicsNET was additionally used to link transcripts and proteins [15]. The current version of the network holds 23947 nodes, each coding for a particular protein (a canonical sequence ensemble is used instead of explicitly representing splice variants). Edges between nodes represent pairwise dependencies which were calculated by integrating similarity and functional dependency measures. A metafunction was used for computing the dependency between nodes resulting in a pair-wise weight matrix, where the weight defines the strength of a dependency. The measures entering the metafunction include each node's tissue specific reference gene expression, conjoint regulation on the level of transcription factors as well as miRNAs, assignment to functional ontologies, subcellular localization, conjoint pathways, as well as protein interaction information. Data sources used for computing the dependency measures included the Gene Expression Omnibus Human Body Map for describing tissue specific gene expression, the MicroCosm database organizing miRNA-target relations, Gene Ontology data on molecular processes and function, PANTHER, KEGG, OPHID, and IntAct database for retrieving protein-protein interactions, complemented by experimentally derived as well as predicted joint transcription factor regulation and subcellular location information.

We used omicsNET in order to identify dependencies between transcripts and proteins thus showing edge weights of two or above (where the edge weights scaled in-between -1.8 and 5.4, where a value of 5.4 represents maximum dependency of a given pair). Based on functional analysis of the given transcriptomics and proteomics features we specifically focused on the blood coagulation cascade.

Additionally, the shortest paths on the omicsNET protein interaction network were calculated between all members of the transcriptomics dataset, the proteomics dataset, as well as between all transcripts and all proteins in both datasets.

Transcription factors

The MAPPER (Multi-genome Analysis of Positions and Patterns of Elements of Regulation) database was used to identify potential direct relationships between transcription factors in the transcriptomics dataset and target genes in the proteomics dataset. MAPPER is a database holding information on putative transcription factor binding sites in the regulatory regions of genes in various species [27].

Kidney tissue expression

Data on immunohistochemical staining in renal tissue were retrieved from the Human Protein Atlas. This source provides a collection of expression and localization data of proteins in normal human tissues, cancer cells and cell lines based on immunohistochemistry and immunofluorescence confocal microscopy images [28]. Data are represented in a semi-quantitative measure with four staining intensities, namely “negative”, “weak”, “moderate”, or “strong”. Staining intensities in the glomerular and the tubular compartments were retrieved from the Human Protein Atlas.

In order to determine mRNA expression levels in kidney tissue, counts of expressed sequence tags were extracted from UniGene EST profiles which show gene expression patterns inferred from EST counts and cDNA library sources. For each tissue and gene, the expression intensity is specified as the occurrence of respective ESTs compared to the total number of reported ESTs in this tissue [29].

RESULTS

Differentially expressed genes and proteins

The transcriptomics dataset consisted of 697 differentially regulated genes, among which 327 showed an upregulation in the diseased state, 355 genes were downregulated, and 15 genes were found to be upregulated in one dataset and downregulated in another dataset. In the 192 urine samples 37 proteins were found in different concentrations when comparing the diseased state and controls.

The genes of four out of the 37 proteins identified as relevant in urine were also differentially expressed in the transcriptomics dataset. The features identified include the collagen, type XV, alpha 1 (COL15A1), and uromodulin (UMOD), as well as the prostaglandin D2 synthase 21kDa (PTGDS) and the apolipoprotein A-I (APOA1) (Table 1).

Symbol	Gene Name	Transcript	Protein
COL15A1	collagen, type XV, alpha 1	up	down
UMOD	uromodulin	up	up
PTGDS	prostaglandin D2 synthase 21kDa	down	up
APOA1	apolipoprotein A-I	down	up

Table 1: Direct overlap of differentially abundant omics features. The table holds gene symbol and name of features being affected at the transcript or protein level, furthermore providing the direct of regulation when comparing diseased and control samples.

Functional overlap

The PANTHER Classification System was used in order to identify enriched biological processes as found on the level of deregulated genes and proteins. Here not the direct feature overlap is determined, but the involvement of transcriptomics and proteomics features in the same pathways and processes. Overall, the biological process of “protein metabolism and modification” was identified as the most significantly enriched, with 153 transcripts assigned to this category but not holding features from proteomics. In contrast, five proteins could be assigned to the biological category “blood circulation and gas exchange” resulting in a p-value smaller than 0.01, without identifying a feature from transcriptomics in this particular functional group.

The four categories that were found to be enriched in both the transcriptomics and proteomics dataset were “cell structure”, “cell structure and motility”, “cell adhesion”, and “immunity and defense”, as listed in Table 2.

Biological Process	# of members total	# of transcripts	p-value	# of proteins	p-value
Protein metabolism and modification	3040	153	< 0.001	-	-
Blood circulation and gas exchange	89	-	-	5	< 0.001
Cell structure and motility	1148	78	< 0.001	8	0.0042
Developmental processes	2152	116	< 0.001	-	-
Immunity and defense	1318	80	< 0.001	9	0.0017
Protein modification	1157	70	< 0.001	-	-
Signal transduction	3406	147	< 0.001	-	-
Cell structure	687	48	< 0.001	8	< 0.001

Cell motility	352	31	< 0.001	-	-
Intracellular protein traffic	1008	57	< 0.001	-	-
Cell cycle	1009	57	< 0.001	-	-
Cell adhesion	622	41	< 0.001	5	0.0478
Cell communication	1213	66	< 0.001	-	-
Intracellular signaling cascade	871	49	< 0.001	-	-
Mesoderm development	551	36	< 0.001	-	-
Mitosis	382	28	< 0.001	-	-
Ectoderm development	692	40	0.0011	-	-
Protein phosphorylation	660	39	0.0011	-	-
Blood clotting	92	12	0.0015	-	-
Cell proliferation and differentiation	1028	50	0.0016	-	-
Cell cycle control	418	28	0.002	-	-
Neurogenesis	587	35	0.0028	-	-
Homeostasis	196	16	0.0034	-	-
Interferon-mediated immunity	63	9	0.0095	-	-
Angiogenesis	54	8	0.0255	-	-
Chromosome segregation	121	12	0.0272	-	-
Apoptosis	531	27	0.0445	-	-

Table 2: PANTHER biological processes overlap. The table lists biological processes identified as relevant on the basis of given transcriptomics and proteomics data sets. Given is the name of the process, the total number of members in the respective process, the number of features involved as found in transcriptomics and proteomics, as well as the p-values regarding the significance of enrichment. Where no p-value is provided the enrichment is not significant for the particular data set. Processes given in bold are significantly enriched by both, transcriptomics and proteomics features.

Joint pathway analysis

Three pathways could be identified as significantly enriched in deregulated transcripts as well as proteins using the KEGG pathway database as repository. Thirteen transcripts and five proteins could be assigned to the “extracellular matrix (ECM)-receptor interaction pathway”, with 18 transcripts and five proteins belonging to the “focal adhesion” pathway (Table 3).

Pathway	# of members total	# of transcripts	p-value	# of proteins	p-value
Cell Communication	136	-	-	6	< 0.001
ECM-receptor interaction	88	13	< 0.001	5	< 0.001
p53 signaling pathway	68	10	0.01	-	-
Complement and coagulation cascades	69	10	0.01	4	< 0.001
Tight junction	132	16	0.01	-	-
Regulation of actin cytoskeleton	214	20	0.02	-	-
Focal adhesion	199	18	0.05	5	< 0.001

Table 3: KEGG pathways overlap. The table lists pathway names, total number of members in the respective pathway, number of involved features from transcriptomics and proteomics, as well as significance of enrichment as found for the respective number of features. Pathways given in bold are enriched by both, transcriptomics and proteomics features.

In addition the “complement and coagulation cascade” was enriched in deregulated features with ten transcripts and four proteins being members of this specific pathway. The coagulation pathway is schematically given in Figure 1.

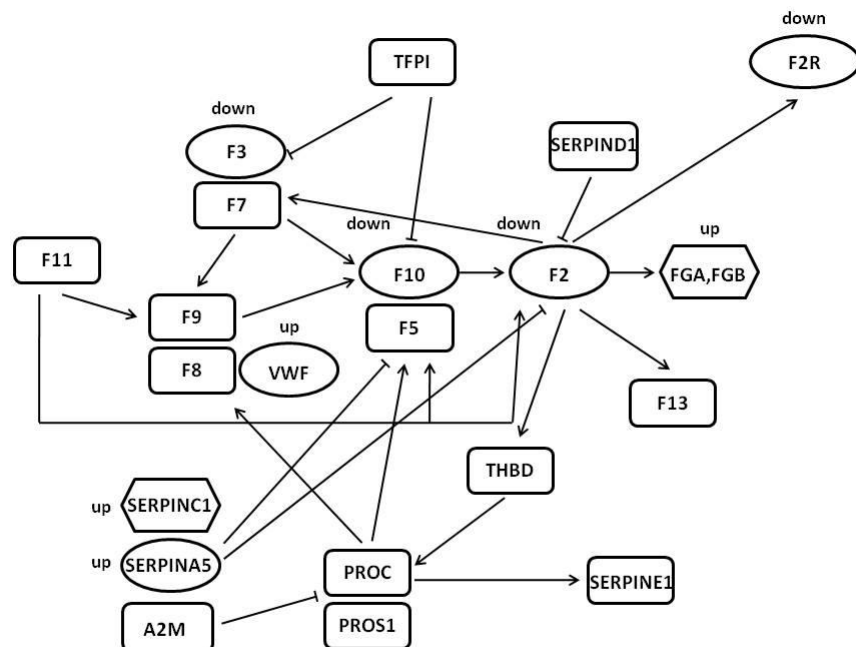


Figure 1: KEGG coagulation pathway. The figure displays a schematic representation of the coagulation pathway as provided by the KEGG pathway database. Transcripts are depicted as oval nodes whereas proteins are given as hexagons.

Protein dependency graph analysis

We identified 65 strong dependencies between features of transcriptomics and proteomics in omicsNET. These dependencies were formed between 21 proteins, 21 transcripts and two features, namely APOA1 and COL15A1, which were found in both omics profiles (figure 2). A large fraction of features was involved in blood coagulation with another highly interconnected subgraph consisting of cell structure and cell adhesion molecules, mainly collagens along with fibronectin 1 (FN1), laminin gamma 3 (LAMC3), and the thrombospondins 1 and 3 (THBS1 and THBS3).

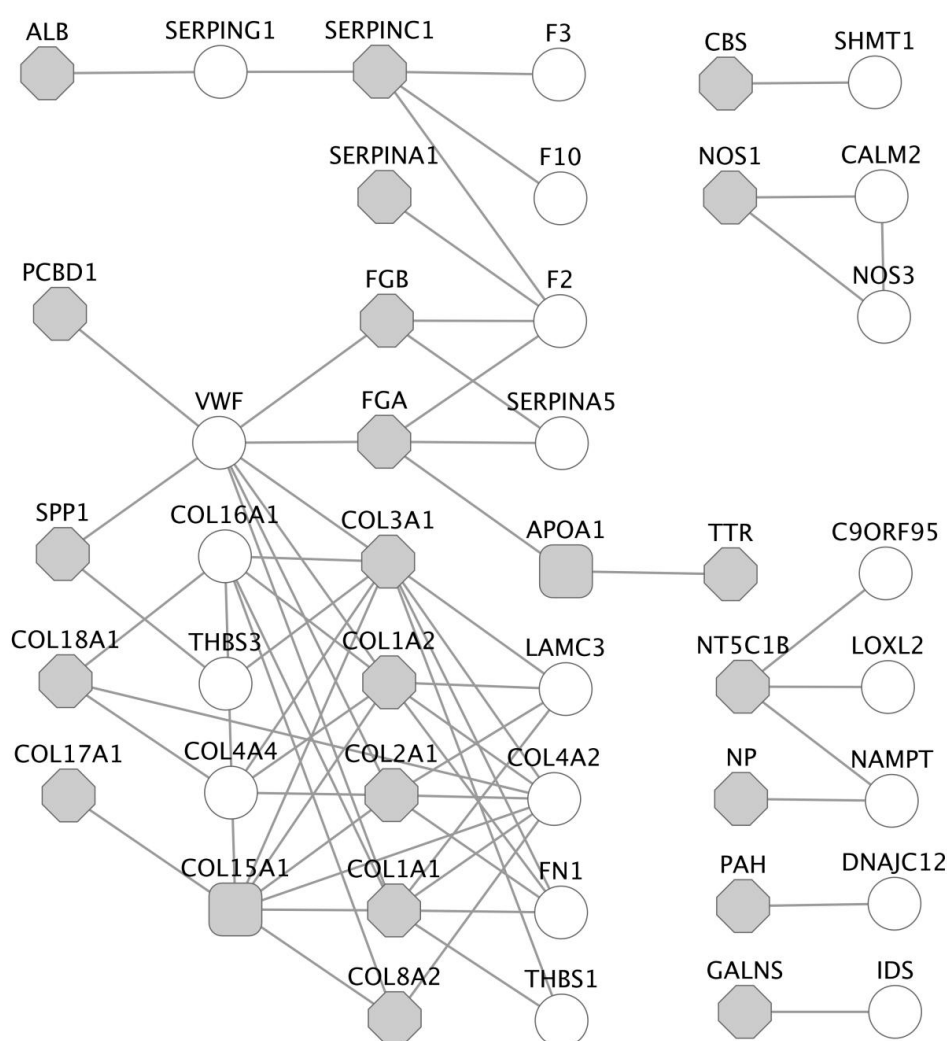


Figure 2: OmicsNET dependencies between transcriptomics and proteomics. The figure displays strong dependencies between transcripts and proteins as derived from omicsNET. Grey nodes represent identified proteins while white nodes represent identified transcripts. The two square nodes represent APOA1 and COL15A1 found with differential abundance in both omics profiles.

Features involved in the blood coagulation cascade according to gene ontology terms were separately analyzed in omicsNET at different cutoff values of computed dependencies (figure 3). 32 edges could be extracted connecting 15 nodes (10 transcripts and 5 proteins) using an omicsNET edge weight of 1. The proteins fibrinogen alpha chain (FGA) and fibrinogen beta chain (FGB), as well as the two serine peptidase inhibitors clade A member 1 (SERPINA1) and clade C member 1 (SERPINC1) all had seven connections to deregulated transcripts. When using an edge weight cutoff of two or above, twelve of the fifteen molecules remained in the network having at least one edge. In total thirteen edge weights had values of two and above with the serine peptidase inhibitor clade C1 (SERPINC1) showing four edges to the coagulation factors II (F2, thrombin), III (F3, thromboplastin), and X (F10) as well as SERPING1.

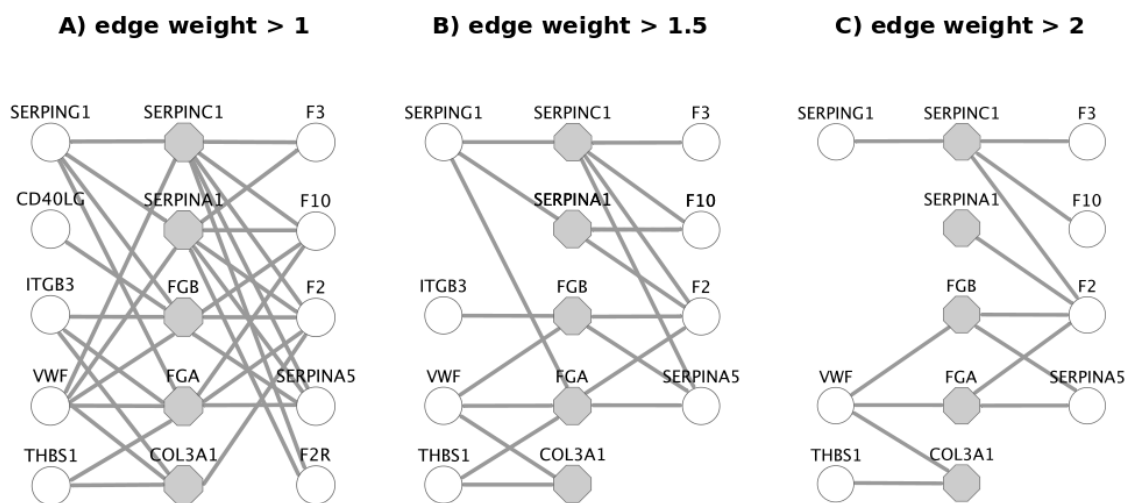


Figure 3: OmicsNET subgraphs of members involved in blood coagulation. The figure shows dependencies as derived from omicsNET analyzing transcripts and proteins involved in the blood coagulation cascade. Figure 3A (edge weight cutoff 1.0) holds 15 nodes and 32 edges, the corresponding number of nodes and edges for a cutoff of 1.5 is 13/19 (3B), and for a cutoff of 2.0 the numbers are 12/13 (3C).

The distribution of shortest paths between members of the transcriptomics list and between members of the transcriptomics and proteomics list were found to be equivalent, again indicating a strong functional link between these two feature lists (figure 4). The distribution of shortest paths was shifted to even shorter values for the

proteomics dataset, partly caused by functional paralogs prevalent in the proteomics dataset.

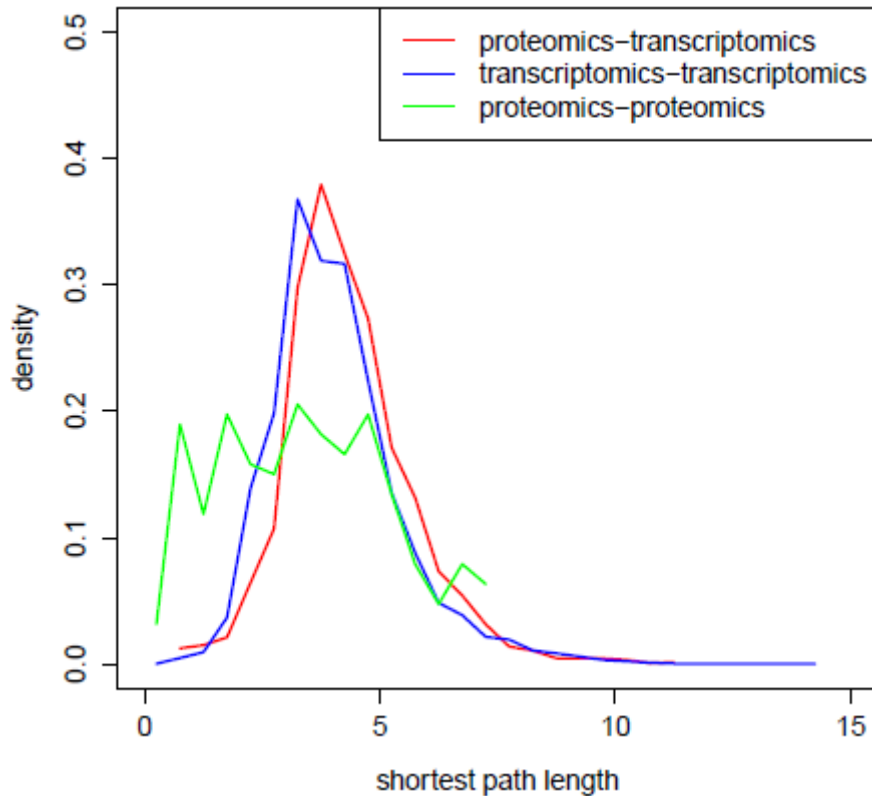


Figure 4: OmicsNET shortest paths distribution. The figure shows the distribution of shortest paths between members of the transcriptomics and the proteomics list as well as between members of the transcriptomics and the proteomics list. Given is the number of nodes connecting two given features (shortest path length) and the number of paths at a certain length represented as density.

Direct edges between transcripts and proteins

Transcription factor binding sites of the factors SP3, IRF9, STAT1, and VDR were identified in the open reading frame regulatory regions of the 37 features from the proteomics dataset. SP3 and ISGF3G were upregulated on the mRNA level whereas VDR and STAT1 showed downregulation. Thirteen proteins had on the gene level a binding site for at least one of the four transcription factors listed above. COL2A1 had binding sites for IRF9 and SP3, A1BG showed binding sites for IRF9 and STAT1, and VGF had binding sites for SP3 and STAT1.

Tissue specific protein expression

Protein expressions levels in renal tissues were determined using data from the publicly available Human Protein Atlas for the proteins given in our dataset. Data were available for 25 out of the 37 proteins of the proteomics set. About 75% of the proteins did show at least weak staining in the tubular compartment, whereas 40% of the proteins did show positive staining in the glomerular compartment (figure 5). Four proteins were neither positive in the tubular nor in the glomerular compartment following the immunohistochemical staining. On the other hand uromodulin (UMOD) and the prostaglandin D2 synthase 21kDa (PTGDS), two proteins also deregulated on the mRNA level, were among the proteins showing the strongest staining in the tubular compartment. The other two proteins also found in the transcriptomics dataset, namely the apolipoprotein A1 (APOA1) and the collagen type XV alpha 1 (COL15A1), did show weak to moderate staining in both, the tubular and the glomerular compartment.

Gene Symbol	G	T	staining intensity
PGRMC1			strong
SPP1			moderate
PTGDS			weak
TTR			negative
UMOD			not available
ALB			
B2M			
FGA			
ORM1			
AHSG			
FXD2			
SERPINC1			
APOA1			
COL15A1			
COL3A1			
CSTB			
FGB			
PIGR			
SERPINA1			
COL1A1			
CD99			
A1BG			
COL18A1			
COL2A1			
PCSK1N			
Gene Symbol	G	T	
COL1A2	X	X	
COL8A2	X	X	
COL17A1	X	X	
HBA1	X	X	
HBA2	X	X	
HBB	X	X	
IGL@	X	X	
IGLC2	X	X	
IGLV2-14	X	X	
PSORS1C2	X	X	
VGF	X	X	
ZNF653	X	X	

Figure 5: Protein tissue staining. The figure displays semi-quantitative tissue staining results in the glomerular (G) and tubular (T) compartment for 25 out of the 37 proteins found in proteomics and also present in the Human Protein Atlas. Staining intensity values range from negative, weak, moderate, and strong as indicates by the different grey shadings. No staining results were available for 12 proteins indicated by "X".

DISCUSSION AND CONCLUSION

Large scale, public domain omics data repositories have been established covering various cellular phenotypes. These data sets allow the analysis of a particular cellular state separately on e.g. the transcript or protein level. However, as these repositories grow the chance of identifying multiple omics levels covering a given analysis question continuously increases.

Joint analysis of transcriptomics and proteomics profiles appears obvious following the general assumption that a change on the mRNA level leads to a change on the protein level. Various studies demonstrate the overall correctness of this assumption but still showing a significant deviation of transcriptome and proteome profiles measured for the very same cellular system. Next to intrinsic biological effects as e.g. variable life time of mRNA and encoded protein following posttranslational modification also other effects are relevant, as e.g. imposed by experimental biases found for both, microarrays as well as proteomics procedures.

This paper analyzed transcriptomics and proteomics profiles derived in the context of chronic kidney disease. Available gene expression data from kidney biopsies resulted in 697 differentially regulated features, proteomics profiles from urine showed 37 proteins as being differentially abundant when comparing chronic kidney disease and healthy reference. This large difference is certainly driven by the different sample matrix analyzed, as even in the presence of chronic kidney disease only a limited number of proteins is released into the urine.

The overlap of transcriptomics and proteomics features is low and ambivalent. The disease associated feature UMOD is found in both data sets as upregulated, whereas three other jointly found features differ in their regulation. PTGDS is mainly expressed in heart and brain tissue and its urinary excretion is closely associated with vascular injury and the following damage of renal interstitial regions [30]. Thus, high PTGDS concentration in urine is not necessarily a consequence of elevated mRNA expression levels in kidney tissue but rather a consequence of damaged vessels and an increased permeability of the kidney filtration barrier.

As reported by Attmann and colleagues, diabetic nephropathy is accompanied with dyslipidemia and, in contrast to most of the other apolipoproteins, decreased plasma levels of APOA1 [31]. These decreased levels in plasma may be due to increased

levels in urine because of a reduced re-absorption from tubules and to low expression levels in kidney tissue.

The deposition of collagens in the extracellular matrix is reported as associated with renal fibrosis [32]. Hagg and colleagues detected high concentration of COL15A1 in kidney biopsies taken from patients suffering from glomerular diseases with interstitial fibrosis [33]. The accumulation of COL15A1 in kidney tissue may lead to a decreased COL15A1 excretion and thus, to decreased COL15A1 levels in urine.

Based on these results the correlation between mRNA and protein abundance on the mere feature level appears limited. In the given case the different sample matrices used for profiling may contribute to this finding. Altered protein abundance resulting from differential gene expression in kidney tissue will not necessarily be reflected by a change of the very same proteins in urine. High concentration of proteins in urine can be caused by an increased permeability of the glomerular filtration barrier for macromolecules. During the progression of chronic kidney disease, a rearrangement of the actin cytoskeleton of glomerular epithelial cells can be observed subsequently leading to proteinuria.

Nevertheless, differential gene expression in chronic kidney disease reflects changes in particular molecular processes and pathways. In turn, features being players in these pathophysiological processes may well be found as proteins in urine. For testing this hypothesis we used directed as well as undirected protein interaction networks for joint analysis of transcriptomics and proteomics features. Directed interaction graphs were drawn from KEGG and PANTHER, and transcriptomics as well as proteomics features were mapped on these graphs. The subsequent analysis focused on the question if dedicated pathways were found to be significantly populated by transcriptomics or proteomics features, or both. Numerous pathways were found affected on the basis of the transcriptomics features, and in PANTHER the processes 'Cell structure and motility', 'Immunity and defense', 'Cell structure' as well as 'Cell adhesion' were significantly populated by features from both data sources. For KEGG the pathways 'ECM-receptor interaction', 'Complement and coagulation cascade' and 'Focal adhesion' were identified on the basis of both sources. Most of the pathways and biological processes reported in the context of CKD are associated with inflammation, cell structure, and cell adhesion. Perco and colleagues presented a list of 11 protein markers of CKD and although the direct overlap between this list and the protein dataset derived from HUPDB consists of only two features (COL3A1, PTGDS), the two

important biological processes 'immunity and defense' and 'cell structure and motility' were found to be enriched in both of the lists [34].

Another functional category found to be overpopulated by transcriptomics and proteomics features is the coagulation pathway. It is frequently reported that patients with CKD exhibit features of a hypercoagulable state which is also a main contributor to subsequent cardiovascular diseases. Eight features of the coagulation pathway seem to be deregulated in case of CKD, including the platelet-vessel wall mediator von Willebrand factor (VWF) and the two plasma protease inhibitors SERPINC1 and SERPINA5. The mRNA expression of some of the coagulation factors (F2, F3, F10) is downregulated which may reflect a regulatory mechanism of the cell to counterbalance high concentrations of pro-coagulation factors in the surrounding kidney tissue.

Mapping omics features on KEGG or PANTHER has its limitations of coverage. Of the 697 features resulting from transcriptomics 233 were found in KEGG and 681 in PANTHER; the corresponding numbers for the 37 proteins are 14 and 35. For overcoming these limitations we used the undirected interaction network omicsNET which covers all presently annotated protein coding genes. Strong edges with edge weight over 2 were identified between 22 members from the transcriptomics and 25 members from proteomics list. Features could be mainly assigned to the functional classes of 'blood clotting', 'cell structure', 'cell adhesion', and 'immunity and defense'. Twelve members of the network spanned by the 22 transcripts and 25 proteins could be assigned to the GO term 'coagulation' and thus, the resulting subgraph represents an extended interaction network of factors involved in the process of coagulation when compared to the coagulation pathway from the KEGG database. When slightly decreasing the cutoff for edge weights, fifteen members of the coagulation cascade could be identified as strongly interconnected. These results indicate the crucial role of hypercoagulability in CKD.

Further validation of the link of the proteomics data set measured in urine and protein abundance given in kidney compartments was performed on the protein level. The glomerular and tubular abundance of 25 out of the 37 proteins identified in proteomics were available as immunohistochemical staining from the Human Protein Atlas. Six out of the 25 were found in substantial concentration in either glomeruli or tubuli, 15 were found as weak or moderate, and only four were not identified in kidney tissue at all,

namely A1BG, COL18A1, COL2A1 and PCSK1N. Following the UniGene EST profiles however, high mRNA levels of COL18A1, COL2A1, and PCSK1N can be found in kidney tissues. ESTs of A1BG mRNA could not be detected in kidney tissues so far.

Integrated analysis of omics profiles provides only moderate add-on information when solely aimed at identifying and subsequently correlating joint features. This fact already becomes evident within omics domains, as exemplified in meta-analyses of e.g. gene expression profiles on cancer and becomes even clearer when spanning different omics levels e.g. involving transcriptomics and proteomics [15,34].

Mapping of heterogeneous omics profiles on protein interaction networks provides an alternative for joint omics feature analysis. From such a joint analysis view pathways and processes characteristic for the phenotype under analysis may become evident.

ACKNOWLEDGEMENTS

Financial support for this study was obtained from the Austrian Research Promotion Agency (Project Number 814.289) and the European Union Framework 7 project SysKid (project number 241544).

The authors declare that they have no commercial conflicts of interest.

REFERENCES

1. Butte A. The use and analysis of microarray data. *Nat Rev Drug Discov* **2002**; 1:951-960.
2. Hanash S. Disease proteomics. *Nature* **2003**; 422:226-232.
3. Perco P, Rapberger R, Siehs C, et al. Transforming omics data into context: bioinformatics on genomics and proteomics raw data. *Electrophoresis* **2006**; 27:2659-2675.
4. Barrett T, Troup DB, Wilhite SE, et al. NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Res* **2009**; 37:D885-890.
5. Gygi SP, Rochon Y, Franza BR, Aebersold R. Correlation between protein and mRNA abundance in yeast. *Mol. Cell. Biol* **1999**; 19:1720-1730.
6. Kadota K, Tominaga D, Asai R, Takahashi K. Correlation Analysis of mRNA and Protein Abundances in Human Tissues. *Genome Lett.* **2003**; 2:139-148.
7. Lu P, Vogel C, Wang R, Yao X, Marcotte EM. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat Biotechnol* **2006**; 25:117-124.
8. Shankavaram UT, Reinhold WC, Nishizuka S, et al. Transcript and protein expression profiles of the NCI-60 cancer cell panel: an integromic microarray study. *Molecular Cancer Therapeutics* **2007**; 6:820-832.
9. Rogers S, Girolami M, Kolch W, et al. Investigating the correspondence between transcriptomic and proteomic expression profiles using coupled cluster models. *Bioinformatics* **2008**; 24:2894-2900.
10. Cox J, Mann M. Is Proteomics the New Genomics? *Cell* **2007**; 130:395-398.
11. Greenbaum D, Colangelo C, Williams K, Gerstein M. Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome Biol* **2003**; 4:117.
12. Greenbaum D. Analysis of mRNA expression and protein abundance data: an approach for the comparison of the enrichment of features in the cellular population of proteins and transcripts. *Bioinformatics* **2002**; 18:585-596.
13. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Research* **2009**; 38:D355-D360.
14. Brown KR. Online Predicted Human Interaction Database. *Bioinformatics* **2005**; 21:2076-2082.
15. Bernthaler A, Mühlberger I, Fechete R, Perco P, Lukas A, Mayer B. A dependency graph approach for the analysis of differential gene expression profiles. *Mol. BioSyst.* **2009**; 5:1720.

16. Hallan SI. International Comparison of the Relationship of Chronic Kidney Disease Prevalence and ESRD Risk. *Journal of the American Society of Nephrology* **2006**; 17:2275-2284.
17. Rudnicki M, Eder S, Perco P, et al. Gene expression profiles of human proximal tubular epithelial cells in proteinuric nephropathies. *Kidney Int* **2006**; 71:325-335.
18. Rudnicki M, Perco P, Enrich J, 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.
19. Rossing K, Mischak H, Dakna M, et al., on behalf of the PREDICTIONS Network. Urinary Proteomics in Diabetes and CKD. *Journal of the American Society of Nephrology* **2008**; 19:1283-1290.
20. Schmid H, Boucherot A, Yasuda Y, et al. Modular activation of nuclear factor-kappaB transcriptional programs in human diabetic nephropathy. *Diabetes* **2006**; 55:2993-3003.
21. Baelde HJ, Eikmans M, Doran PP, Lappin DWP, de Heer E, Bruijn JA. Gene expression profiling in glomeruli from human kidneys with diabetic nephropathy. *American Journal of Kidney Diseases* **2004**; 43:636-650.
22. Coon JJ, Zürbig P, Dakna M, et al. CE-MS analysis of the human urinary proteome for biomarker discovery and disease diagnostics. *Proteomics Clin Appl* **2008**; 2:964.
23. Bairoch A. Swiss-Prot: Juggling between evolution and stability. *Briefings in Bioinformatics* **2004**; 5:39-55.
24. Mi H, Guo N, Kejariwal A, Thomas PD. PANTHER version 6: protein sequence and function evolution data with expanded representation of biological pathways. *Nucleic Acids Research* **2007**; 35:D247-D252.
25. Dennis G, Sherman BT, Hosack DA, et al. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol* **2003**; 4:P3.
26. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **2008**; 4:44-57.
27. Marinescu VD, Kohane IS, Riva A. The MAPPER database: a multi-genome catalog of putative transcription factor binding sites. *Nucleic Acids Res* **2005**; 33:D91-97.
28. Berglund L, Björling E, Oksvold P, et al. A gene-centric Human Protein Atlas for expression profiles based on antibodies. *Mol. Cell Proteomics* **2008**; 7:2019-2027.
29. Boguski MS, Lowe TM, Tolstoshev CM. dbEST--database for "expressed sequence tags." *Nat. Genet* **1993**; 4:332-333.
30. Yoshikawa R, Wada J, Seiki K, et al. Urinary PGDS levels are associated with vascular injury in type 2 diabetes patients. *Diabetes Res. Clin. Pract* **2007**; 76:358-367.

31. Attman PO, Knight-Gibson C, Tavella M, Samuelsson O, Alaupovic P. The compositional abnormalities of lipoproteins in diabetic renal failure. *Nephrol. Dial. Transplant* **1998**; 13:2833-2841.
32. Alexakis C, Maxwell P, Bou-Gharios G. Organ-specific collagen expression: implications for renal disease. *Nephron Exp. Nephrol* **2006**; 102:e71-75.
33. Hägg PM, Hägg PO, Peltonen S, Autio-Harmainen H, Pihlajaniemi T. Location of type XV collagen in human tissues and its accumulation in the interstitial matrix of the fibrotic kidney. *Am. J. Pathol* **1997**; 150:2075-2086.
34. Perco P, Pleban C, Kainz A, et al. Protein biomarkers associated with acute renal failure and chronic kidney disease. *Eur. J. Clin. Invest* **2006**; 36:753-763.

2.4.1 The Thesis Author's Contribution

The thesis author was primarily responsible for the selection of the transcriptomics and proteomics datasets, as well as for the functional analyses. Furthermore, the author contributed to the study design and the interpretation of the results.

In detail, the following contributions are due to the thesis author's efforts:

- Contributions to selection of transcriptomics and proteomics datasets, as well as of appropriate bioinformatics tools
- Retrieval of publicly available transcriptomics datasets on diabetic nephropathy [20,21] and non-diabetic nephropathies [17] from the Gene Expression Omnibus database and respective publications
- Extraction of proteins associated with chronic kidney diseases from the Human Urinary Proteome Database
- Accomplishment of the pathway enrichment analysis using the PANTHER classification tool
- Extraction of kidney specific protein tissue expression from the Human Protein Atlas
- Contributions to the interpretation of results from the functional interaction analyses
- Visualization of the pathway and networks
- Provision of bioinformatics-specific methods and results sections to the manuscript draft

Integrative bioinformatics analysis of proteins associated with the cardiorenal syndrome

Irmgard Mühlberger¹, Konrad Moenks², Andreas Bernthaler³,
Christine Jandrasits¹, Bernd Mayer¹, Gert Mayer², Rainer Oberbauer^{3,4}, and
Paul Perco^{1,4 *}

¹ emergentec biodevelopment GmbH, Gersthofer Strasse 29-31, 1180 Vienna, Austria

² Medical University of Innsbruck, Department of Internal Medicine IV, Anichstrasse 35, 6020 Innsbruck, Austria

³ KH Elisabethinen Linz, Fadingerstrasse 1, 4020 Linz, Austria

⁴ Medical University of Vienna, Department of Internal Medicine III, Waehringer Guertel 18-20, 1090 Vienna, Austria

* Corresponding author:

Dr. Paul Perco
emergentec biodevelopment GmbH
Gersthofer Strasse 29-31
1180 Vienna, Austria
phone: +43-1-4034966
fax: +43-1-4034966-19
e-mail: paul.perco@emergentec.com

ABSTRACT

The cardiorenal syndrome refers to the coexistence of kidney and cardiovascular disease, where cardiovascular events are the most common cause of death in patients with chronic kidney disease. Both, cardiovascular as well as kidney diseases have been extensively analyzed on a molecular level, resulting in molecular features and associated processes indicating a cross-talk of the two disease etiologies on a pathophysiological level.

In order to gain a comprehensive picture of molecular factors contributing to the bidirectional interplay between kidney and cardiovascular system, we mined the scientific literature for molecular features reported as associated with the cardiorenal syndrome, resulting in 280 unique genes/proteins. These features were then analyzed on the level of molecular processes and pathways utilizing various types of protein interaction networks.

Next to well established molecular features associated with the renin-angiotensin system numerous proteins involved in signal transduction and cell communication were found, involving specific molecular functions covering receptor binding with natriuretic peptide receptor and ligands as well known example. An integrated analysis of all identified features pinpointed a protein interaction network involving mediators of hemodynamic change and an accumulation of features associated with the endothelin signaling and VEGF signaling pathway. Some of these features may function as novel therapeutic targets.

INTRODUCTION

The risk of developing cardiovascular disease (CVD) is dramatically increased in patients with chronic kidney diseases (CKD). Mortality as a consequence of cardiovascular events is 10 to 30 times higher in patients on dialysis treatment than in the general population [1]. Due to this recognition of CVD as the leading cause of morbidity and mortality in patients with reduced kidney function, a growing body of literature has become available regarding this link of CKD and CVD, termed as cardiorenal syndrome (CRS).

CRS can be classified into five subtypes depending on the origin of damage (either the cardiovascular system or the kidney) and the course of disease (either acute or chronic) [2,3]. Major mechanisms leading to CRS1 and CRS2 (acute and chronic cardio-renal syndrome) include hemodynamically mediated damage, hormonal factors, immune mediated damage, low cardiac output, endothelial dysfunction, and chronic hypoperfusion. Hallmarks of kidney dysfunction leading to CRS3 and CRS4 (acute and chronic reno-cardiac syndrome) on the other hand are volume expansion, drop of the glomerular filtration rate, humoral signaling, anemia, uremic toxins, and inflammation. The fifth subtype of the cardiorenal syndrome (CRS5) describes the secondary cardio-renal syndrome which refers to systemic diseases such as diabetes that ultimately lead to simultaneous cardiovascular and kidney dysfunction.

The multitude of cardiac risk factors in patients with chronic kidney disease are complex and increase with age, the stage of kidney disease, and the level of proteinuria. Another powerful risk factor is hypertension which goes along with sodium retention, and activation of the renin-angiotensin system. Atherosclerosis results from an impairment of endothelial function which, in turn, is associated with albuminuria. Changes in blood-lipid composition and oxidative stress as a consequence of inflammation due to renal dysfunction also contribute to endothelial dysfunction and subsequent CVD [4].

Management and therapy of the CRS is challenging since drugs in use for the treatment of cardiovascular diseases may go along with impairment of kidney function and vice versa. Examples include diuretics, ionotropes, angiotensin-converting enzyme inhibitors, angiotensin receptor blockers, or natriuretic peptides but treatment decision must be based on a combination of individual patient information and understanding of individual treatment options [5].

Biomarkers of relevance in the context of the CRS mainly hold proteins known either in the field of nephrology or cardiology, for the latter including e.g. the family of natriuretic peptides and troponins, whereas frequently reported renal specific markers include neutrophil gelatinase-associated lipocalin (NGAL), kidney injury molecule 1 (KIM1), Cystatin C, interleukin 18 (IL18), and N-acetyl- β -D-glucosaminidase [6]. Levels of circulating fibroblast growth factor 23 (FGF-23) for example have been shown to be independently associated with left ventricular mass index and left ventricular hypertrophy in patients with CKD [7]. Chung and colleagues described the relationship between activation of matrix metalloproteinase 2 (MMP2) and elastic fiber degeneration, stiffening, medial calcification, and vasomotor dysfunction in macroarterial vasculature of dialyzed CKD patients [8]. Next to these proteins a multitude of other molecular features is mentioned in the literature in the context of the cardiorenal syndrome. Perco et al. reported a list of 31 CVD biomarkers that were extracted from literature and characterized with respect to biological function, gene expression in CKD, and known protein–protein interactions [9].

Literature mining approaches have the potential to reveal such biomarkers, thus providing a more global picture on genes, proteins, and metabolites associated with a specific disease. The biomedical literature can be seen as the condensed result of the combined effort of the scientific community. As such, it represents the primary resource upon which further investigations may be based on. PubMed, for instance, presently holds close to 20 million abstracts. Thus, computational literature mining tools assisting researchers in keeping pace with this ever-growing amount of fast changing information became indispensable [10,11].

In the context of drug discovery, the most prevailing approach is based on concept co-occurrence [12]: Here, a disease profile consisting of the concepts (e.g. drugs, genes, etc.) which are frequently mentioned together with the disease under analysis can be derived via text mining. Likewise, literature based profiles for drugs or genes can be generated. Next to conveniently reaching an overview on biomarkers this information base may additionally be used to gain hints about yet undiscovered dependencies between diseases, drugs, and potential drug targets.

To further enhance text mining efforts, several “controlled vocabularies” (“ontologies”) have been developed to allow a precise definition of the employed concepts [13]. The most popular ones are maintained by the U.S. Library of Medicine, namely the Unified Medical Language System (UMLS) and the Medical Subject Headings (MeSH). Given that the majority of PubMed articles are indexed with MeSH, a fast and accurate

extraction of biomedical concepts has become feasible [14,15]. With the advent of literature mining approaches also in combination with high-throughput Omics experiments, a number of bioinformatics tools and ontologies have been developed for the analysis of resulting large sets of genes or proteins. Analyzing extended sets of biomarker candidates on the level of molecular pathways and processes, represented as protein interaction networks, add another layer of information for the interpretation of molecular feature (biomarker) sets.

A recent review by Lusi and colleagues summarized studies dealing with network analyses in cardiovascular disease [16]. Networks based on prior knowledge, such as existing pathway sources, literature co-citations or other correlation measures as co-expression and sequence similarity were outlined by Ashley et al. [17], who mapped genes being differentially regulated between patients suffering from de-novo atherosclerosis and in-stent restenosis on a co-citation network obtained by literature mining of Medline abstracts. Similar concepts can be followed by utilizing networks derived from physical protein interactions, or networks generated from measuring the response to experimental perturbations. Further approaches include system genetics and detailed analyses at the level of dynamic systems such as flux balance analyses which are often used to characterize enzymatic reactions in dynamic models of metabolism. Some of these approaches, especially highly abstracted network models on the level of phenotypes, managed to predict co-morbidity patterns for myocardial infarction using a 'human disease network' thus closing the gap to clinical applications [18].

Diez et al. presented another application of the network paradigm to reveal the mechanisms of cardiovascular disease, identifying a set of differentially expressed genes separating asymptomatic from symptomatic carotid stenosis patients [19]. Based on these transcriptomics data a correlation network was generated. Furthermore an association network of the differentially regulated genes was derived by mining the literature for gene associations thus resulting in an interaction network combining Omics data and associated features extracted from literature. Sub-networks were identified, characterized by enriched lipid-, immune-, and atherogenesis related pathways and gene ontology terms. On this level of representation the interplay of APOC1 (a gene that is linked to coronary heart disease) became evident. Weiss et al. investigated networks on cardiovascular metabolism pointing out aspects of network structure, namely differences between designed networks in engineering and networks having undergone an evolutionary process [20]. Based on the level of abstraction three types of network on cardiovascular metabolism were proposed: First, on the very

abstract level of nodes and edges, metabolite networks described by using topological characteristics [21,22], second physical, spatially compartmentalized networks including the description of energy fluxes in the network [23,24], and on a third level dynamic networks [25-27].

The present knowledge regarding mechanisms leading to the formation of the CRS suggests a critical role for hemodynamic changes, originating either from the kidney or the cardiovascular system. In the following analysis we used a literature mining approach to extract genes and proteins reported in the context of the cardiorenal syndrome, and analyzed these features on the level of protein interaction networks. Specific focus was laid on secreted proteins being specifically expressed in either renal or vascular tissue with the aim to identify molecular mediators potentially contributing to the cross-talk between the kidney and the cardiovascular system for allowing identification of novel therapeutic targets addressing both systems.

MATERIALS AND METHODS

The general analysis strategy applied in this work is outlined in figure 1. Major components include feature extraction via literature mining, followed by a range of bioinformatics analysis procedures for deciphering characteristics of individual features as well as joint interpretation on the level of protein interaction networks.

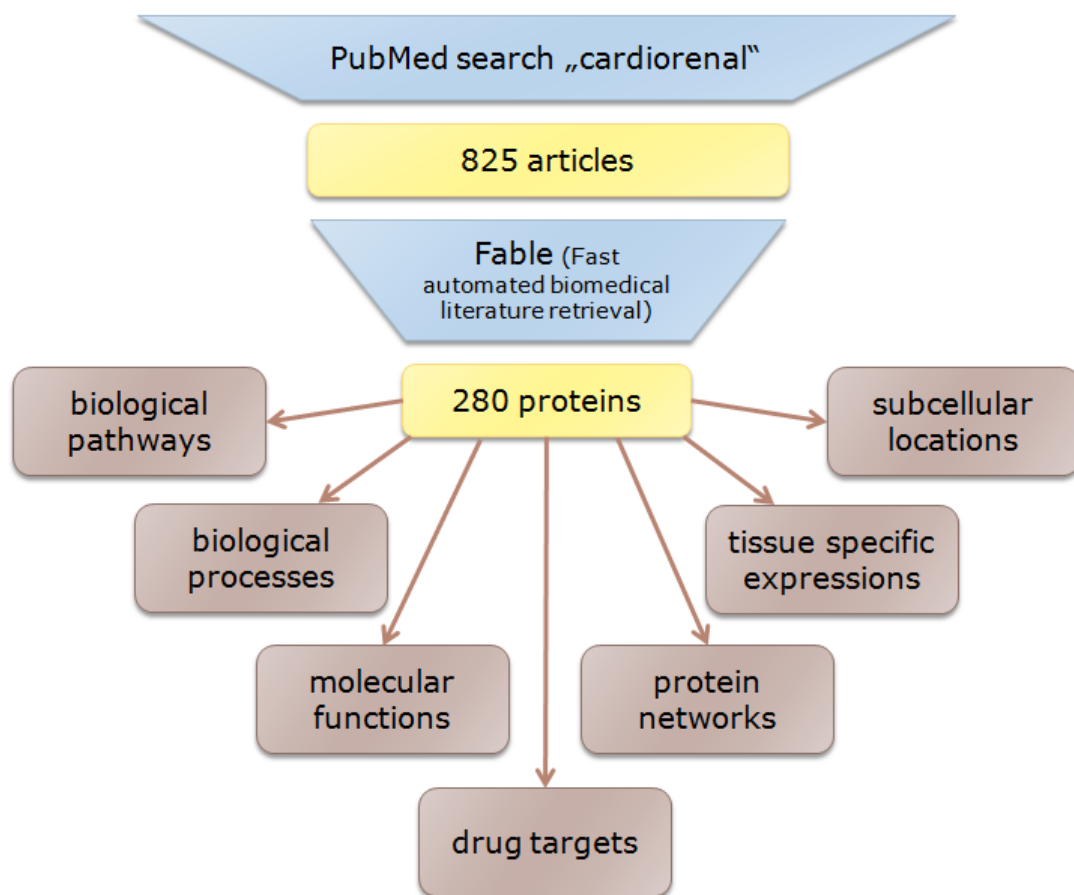


Figure 1: Overview scheme on the analysis workflow: Literature mining was applied for identifying unique proteins associated with CRS. Bioinformatics included feature characterization as well as network analysis.

Literature mining

The strength but also the challenge of biomedical text mining relies on the fact that the scientific literature embraces a variety of concepts (genes, drugs, diseases, etc.) which in turn are inter-related in a variety of ways. Thus, carefully designed text mining methods are needed to extract “meaningful” information and reduce the amount of noise present in the final results.

In general, text mining consists of two steps: Information Retrieval (IR) and Information Extraction (IE) [10]. The first consists in identifying documents which are of relevance for a certain research objective (e.g. a PubMed query for “cardiorenal”), whereas the later is used to extract facts from these documents. Named Entity Recognition (NER) can be seen as the most prevalent type of IE used in real world applications, aiming at the identification of biological entities like genes, cell types or drugs.

Even though the concept of NER might appear almost trivial at a first glance, it actually represents a challenging computational problem as the existence of over fifty available tools demonstrates [28]. The key obstacle that needs to be addressed when extracting genes or proteins from free text relies in the term ambiguity present at multiple levels. Some genes are spelled like normal English words (e.g. “WAS” with the NCBI GeneID: 7454) and even a gene with the official Gene Symbol “T” exists (NCBI GeneID: 6862). The same gene may additionally be referred to in various ways due to different naming conventions.

Ultimately, this ambiguities lead to two different types of errors which all methods are confronted with: erratically assuming that a certain gene was mentioned in a paper (false positive) or erratically assuming that it was not mentioned, even though it actually was given (false negative) [29]. Based on the trade-off between these two types of errors, the precision of a method (i.e. how much of the predicted genes were actually mentioned in the document) and its recall (i.e. how much of all actually mentioned genes were also identified as such) are determined.

We chose a method favoring precision over recall for mining genes/proteins in Medline / PubMed abstracts. The Fast Automated Biomedical Literature Extraction (FABLE) tool available at <http://fable.chop.edu> was used in order to fulfill this task. The algorithm basically consists of two steps: First, a statistical classifier was used to train a probabilistic model, which served as basis for gene tagging, i.e. to identify possible occurrences of a gene, taking the textual context into account. Given that such an

occurrence exhibits a sufficient likelihood of actually representing a gene, this occurrence was normalized in a second step to the official Gene Symbol. This normalization step was based on gene synonym lists, which were compared to the predicted occurrence using both exact and relaxed pattern matching procedures. It has been shown that this approach is competitive to alternative methods such as standard information extraction techniques and direct pattern matching both in terms of precision and recall [30,31]. We applied this procedures for all papers retrieved from PubMed associated with “cardiorenal” (PubMed status as of March 2010).

Functional annotation of identified genes/proteins

The list of genes and proteins identified on the basis of the literature mining approach was in a first step annotated using the Stanford Source tool [32]. The set of genes was assigned to biological processes, pathways, and molecular functions using the PANTHER (Protein Analysis THrough Evolutionary Relationships) Classification System [33,34]. Significantly enriched categories were identified using the whole human genome as reference dataset. Biological processes, pathways, and molecular functions showing p-values below 0.0001 were considered as statistically significant in terms of feature enrichment.

The subcellular location of proteins was determined using experimental data provided by SwissProt [35]. For proteins not covered in SwissProt *in-silico* predictions using WoLF PSORT were done [36]. WoLF PSORT computes probabilities based on the protein sequence of a given protein for ten subcellular locations. Subcellular location tags from SwissProt were mapped to the ten locations defined by WoLF PSORT. Only assignments that were either reported in SwissProt or showed a probability value of 1 according to WoLF PSORT were considered for subcellular location enrichment analysis. Based on a reference dataset of 45,008 proteins assigned to one of the WoLF PSORT categories, the significance of enrichment was calculated using the Fisher's exact test. P-values below 0.01 were considered as statistically significant.

Information on tissue specific expression patterns was extracted from NCBI UniGene EST profiles. EST counts of in total 45 tissues were extracted for each gene. Tissue

specific expression patterns for each single tissue for each single gene were calculated based on the normalized transcripts per million counts as provided by UniGene [37].

Network analysis framework

For network analysis we used an extended version of the protein dependency network “omicsNET” as described in Bernthaler et al. [38]. The network is comprised of information from protein-protein interactions, tissue specific reference co-expression, shared pathway information, gene ontology distance, and subcellular co-localization, and was extended by networks generated from shared transcription factor binding sites and shared miRNA target sites. In omicsNET these sources were consolidated into a single human protein reference interaction network, where edges represent pairwise dependencies between proteins.

Protein-protein dependencies were calculated between proteins in the list resulting from the literature mining approach. Furthermore, highly connected subgraphs were identified and functionally annotated. We only considered dependencies with high confidence in the network construction process and focused on genes reported at least twice in the scientific literature in the context of the cardiorenal syndrome in order to reduce the number of false positive assignments.

Identification of drug targets

Drug targets were identified in our set of 280 literature derived proteins using information from DrugBank [39,40]. DrugBank combines information on drugs and their molecular targets and currently contains around 4800 drug entities with more than 1350 FDA-approved small molecule drugs and more than 2500 protein drug targets.

RESULTS AND DISCUSSION

Literature mining

825 papers associated with the term “cardiorenal” were identified in PubMed. In this set of 825 papers 280 genes could be extracted utilizing FABLE, with 132 genes being reported at least twice. The top ranked gene, mentioned in 156 articles, was the aspartyl protease renin (REN), followed by the natriuretic peptide precursor A (NPPA) and angiotensinogen (AGT), with 122 and 64 reports, respectively.

The list of 54 genes mentioned in at least 5 articles along with the term cardiorenal is provided in Table 1 (see supplementary Table 1 for the total list of 280 genes, available at <http://www.sage-hindawi.com/journals/ijn/2011/809378/sup/>). Next to the number of articles, the relative expression levels in the four tissues blood, heart, vascular, and kidney are provided based on data from the UniGene expressed sequence tag counts.

Symbol	Articles	<i>expression in blood (%)</i>	<i>expression in heart (%)</i>	<i>expression in vascular (%)</i>	<i>expression in kidney (%)</i>	max. expression (%)	
REN	156	0	0	0	19,27	39,58	intestine
NPPA	122	88,04	0	0	0	88,04	heart
AGT	64	1,79	18,54	0	5,71	29,74	liver
ADM	55	0,95	1,38	1,09	3,11	15,3	adipose tissue
ACE	39	0,86	2,37	4,09	4,53	15,63	parathyroid
EDN1	39	0	4,12	15,82	2,77	32,68	umbilical cord
NPPB	31	85,93	0	0	1,2	85,93	heart
RAPGEF5	28	0	0	0	0,76	76,62	parathyroid
NOS3	27	3,92	2,69	2,33	2,2	20,32	spleen
EPO	22	0	0	0	0	58,82	prostate
CNP	21	0,85	1,74	3,58	5,4	18,03	brain
TGFB1	20	8,67	0,99	0	1,79	17,67	salivary gland
MME	19	0,26	3,59	0	11,63	12,06	lymph node
PTGS2	19	16,39	0	29,1	0,59	29,1	vascular
INS	18	0	0	0	0	100	pancreas
NPR1	17	0	1,32	2,29	2,83	23,69	mammary gland
NOS2	13	4,23	0	0	0	25,4	pharynx
DDR1	13	0	0,94	0	0,46	20,12	trachea
KNG1	10	0	0	0	33,18	57,18	liver
PLEK	10	11,02	0,34	1,77	0,87	16,81	lymph
NCF1	10	10,88	0	0	0,76	32,38	lymph node

HESX1	10	0	0	0	0	43,18	ovary
FOS	9	19,04	2,09	4,31	0,77	19,04	blood
CALCA	9	0	0	0	0	100	prostate
S100A6	9	1,2	0,87	5,16	1,18	20,08	umbilical cord
NOS1	8	0	0	0	1,68	65,97	muscle
AVP	8	0	0	0	80	80	kidney
RHOA	7	2,5	1,57	2,02	1,72	5,28	cervix
CYBB	7	19,44	0	2,55	3,15	27,68	lymph node
MAPK1	7	1,84	1,35	2,36	1,44	10,94	mouth
AKT1	7	1,14	1,57	0,45	1,51	13,52	salivary gland
ICAM1	7	3,19	0,55	2,39	1,62	15,19	spleen
CALCRL	7	0	2,55	14,85	1,39	25,06	trachea
SERPINE1	7	0,17	0,12	14,5	0,69	27,77	umbilical cord
EDNRA	7	0	6,4	2,21	1,63	10,94	uterus
SHBG	7	0	0	0	0	36,84	eye
RAMP2	7	5,09	0	0	1,85	28,7	thyroid
UTS2	7	0	0	0	3,88	35,92	spleen
OLR1	6	1,23	0	0	2,15	81,05	esophagus
AGTR1	6	0	5,19	0	3,3	19,1	larynx
NFKB1	6	4,69	0,76	0,66	1,62	8,69	nerve
UTS2R	6	0	0	0	0	100	ovary
NR3C2	6	0	0	6,41	7,08	20,74	stomach
EPHB2	6	6,73	0	0	2,85	14,78	umbilical cord
ISYNA1	6	1,49	0,43	0,52	3,31	17,72	umbilical cord
GPR182	5	0	0	0	0	38,67	adrenal gland
COX8A	5	0,77	11,02	1,48	0,98	11,02	heart
CPOX	5	9,24	3,63	0	5,28	11,06	liver
EGFR	5	0	2,2	1,69	2,49	14,89	mouth
COX5A	5	0	0	0	0	100	muscle
CCL2	5	0	0	0	0	100	placenta
PPARG	5	0	1,46	2,52	3,72	12,08	placenta
CYBA	5	2,25	6,82	1,67	3,43	15,46	tonsil
RAMP3	5	7,76	0	0	2,54	21,44	adipose tissue

Table 1: List of identified genes/proteins, number of articles identified for cardiorenal, and relative expression levels based on UniGene EST counts for blood, heart, vascular and kidney, and tissue showing maximum expression of a specific feature.

The top ranked feature in the list of 280 literature derived genes is renin (REN) which is secreted by cells of the juxtaglomerular apparatus of the kidney and plays a key role in

the blood pressure and water balance-regulating renin-angiotensin system (RAS). The connection between CRS and an increased activity of this hormone system was first reported in 1971 [41] and its consequences like renal hypoxia, vasoconstriction, intraglomerular hypertension, glomerulosclerosis, tubulointerstitial fibrosis, and proteinuria continue to be demonstrated in clinical practice. Conservative therapy for blocking the RAS activity is the administration of angiotensin-converting enzyme inhibitors and angiotensin receptor blockers, but recent studies demonstrate the benefit of a combination with direct renin inhibitors [42].

Further genes frequently reported in association with CRS are the components of the natriuretic peptide system (NPS) NPPA and NPPB, as well as their receptors NPR1, NPR2, and NPR3. Functions of the NPS include the counter-regulation of RAS, and it is suggested that its activation provides organ protection in cardiorenal disease, especially in diabetic patients [43].

Functional annotation

According to the PANTHER Classification System, the biological processes of “signal transduction” and “cell communication” were identified as most significantly enriched, with 135 and 136 genes assigned to these categories, respectively. In total, 28 processes showed a p-value > 0.0001 in terms of enrichment, including “blood circulation”, “regulation of vasoconstriction”, and “angiogenesis”. The most significantly enriched molecular functions are “receptor binding” and “protein binding” (Table 2).

Biological Process	No. genes total	No. genes CRS	No. genes CRS expected	P-value
signal transduction	4191	135	57,67	4.55E-25
cell communication	4365	136	60,07	6.84E-24
cell surface receptor linked signal transduction	2235	91	30,76	3.80E-22
immune system process	2628	97	36,16	9.70E-21
blood circulation	210	28	2,89	5.11E-19
regulation of biological process	59	18	0,81	1.01E-18
regulation of vasoconstriction	59	18	0,81	1.01E-18

Molecular Function	No. genes total	No. genes CRS	No. genes CRS expected	P-value
receptor binding	1233	64	16,97	2.46E-20
protein binding	3157	103	43,44	2.71E-18
catalytic activity	5336	128	73,43	1.44E-12
oxidoreductase activity	703	33	9,67	1.21E-09
binding	6751	140	92,9	3.65E-09
kinase activity	695	28	9,56	5.18E-07

Table 2: List of enriched biological processes and molecular functions. Given is the total number of genes assigned to a process/function, the number of genes assigned as derived from literature mining, the number of genes expected from a statistical perspective, and the significance level of enrichment.

The two enriched categories “receptor binding” and “receptor activity” indicate that numerous receptors and ligands are involved in the cardiorenal syndrome. These receptors form the first line of molecules in a number of signaling cascades, which as such is another category enriched in genes associated with the cardiorenal syndrome. We therefore took a closer look at receptor-ligand interactions. We searched for receptors mainly expressed in the cardiovascular system having ligands predominantly secreted by the renal tissue, and vice versa.

The natriuretic peptide receptor NPR3 showed high expression in kidney tissue, whereas the ligands NPPA and NPPB were found to be almost exclusively expressed in the heart. Thus, a deregulation of blood pressure maintenance and extracellular fluid volume by heart derived ligands of the natriuretic peptide system directly affect the kidney and may contribute to the formation of CRS.

Enrichment of the process “regulation of vasoconstriction” reflects the consequences of impaired heart function including a decreased cardiac output, and thus the hypoperfusion of organs. Since glomerular filtration is controlled by blood pressure, hypoperfusion of the kidney leads to the activation of the RAS and subsequent vasoconstriction, which, in turn, causes systemic hypertension and an increased heart preload [2].

22 PANTHER pathways could be identified as significantly enriched in the list of 280 literature derived genes. 28 genes could be assigned to “angiogenesis”, 21 genes to “endothelin mediated signaling”, and 15 genes to the “VEGF signaling pathway” (Table 3).

Pathway	No. genes total	No. genes CRS	No. genes CRS expected	P-value
Angiogenesis	191	28	2,63	4.51E-20
Endothelin signaling pathway	91	21	1,25	3.33E-19
VEGF signaling pathway	75	15	1,03	3.33E-13
Inflammation mediated by chemokine and cytokine signaling pathway	283	24	3,89	2.76E-12
PDGF signaling pathway	159	18	2,19	1.68E-11
T cell activation	102	14	1,4	2.72E-10
Apoptosis signaling pathway	123	15	1,69	3.10E-10

Table 3: List of enriched biological pathways. Given is the total number of genes assigned to a process/function, the number of genes assigned as derived from literature mining for CRS, the number of genes expected from a statistical perspective, and the significance level of enrichment.

The connection between angiogenic processes and cardiovascular disorders is well understood, since decreased cardiac output goes along with decreased organ perfusion, and vascularization is the natural response to diminution of blood supply. Apart from negative effects on organ function due to hypoperfusion, microvascularization is extensively performed at sites of inflammation which explains the role of angiogenesis in diseased kidney tissue. On the other hand, decreased vascularization and loss of capillaries lead to kidney fibrosis. However, deregulation of angiogenesis seems to be crucial for kidney function and a key regulatory mechanism of angiogenic processes is the VEGF signaling pathway [44-46]. A third enriched pathway is the “endothelin signaling pathway” which is known to regulate the renin-angiotensin system thus being a further player in the hemodynamic crosstalk between the kidney and the cardiovascular system.

Following the rationale that features secreted from kidney cells may lead to damage in vessels and vice versa, literature derived proteins were classified in terms of subcellular location. The most significantly enriched compartment was “extracellular, including cell wall” with 81 genes being assigned to this category, whereas “nuclear” was significantly depleted with 48 genes as indicated in Figure 2.

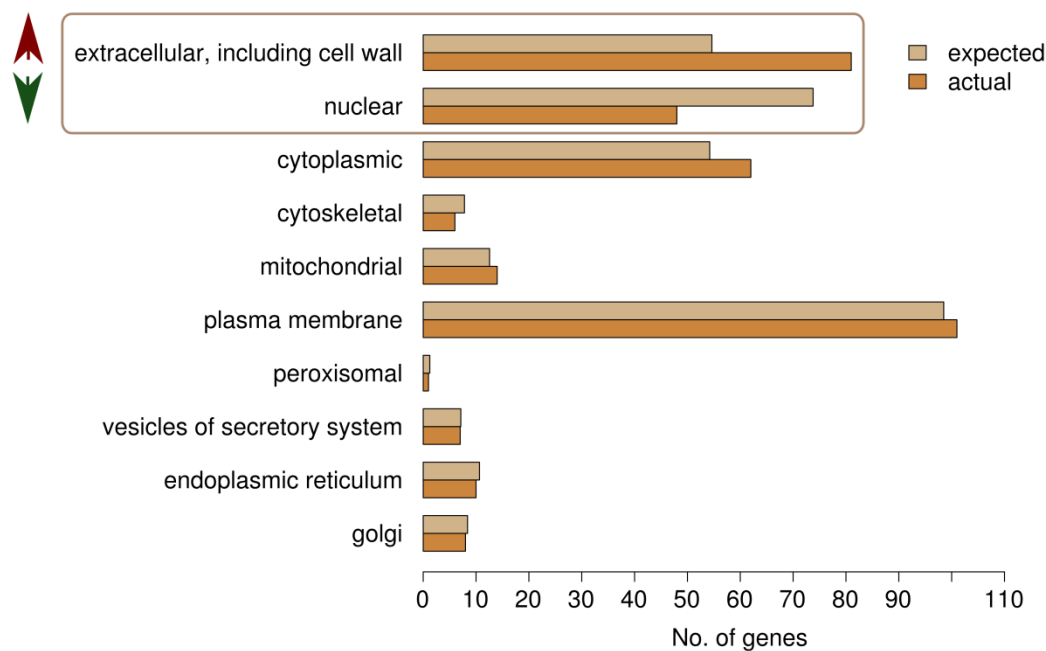


Figure 2: Subcellular location of literature derived proteins. Presented are categories of subcellular location, the expected number of proteins in a particular category using the total set of human proteins, and the actual number of proteins found as being associated with CRS.

The list of 81 secreted genes included components of the renin-angiotensin system (REN, AGT, ACE) and the natriuretic peptide system (NPPA, NPPB), as well as some other regulators of vasoconstriction. Kininogen 1 (KNG1) for example is essential for the assembly of the blood pressure regulating kallikrein-kinin system. Another molecule serving as a vasodilator is the peptide hormone calcitonin-related polypeptide alpha (CALCA).

Network analysis

A subset of 40 proteins out of the list of 132 proteins mentioned in at least two publications in the context of the cardiorenal syndrome formed a highly connected protein interaction network as given in Figure 3. The main components of this protein network are mediators of hemodynamic change. An accumulation of features involved in previously described signaling pathways like the endothelin signaling pathway or the VEGF signaling pathway is evident. Next to these two pathways, a number of members

of the blood pressure regulating kallikrein-kinin system and the renin-angiotensin system are part of this network.

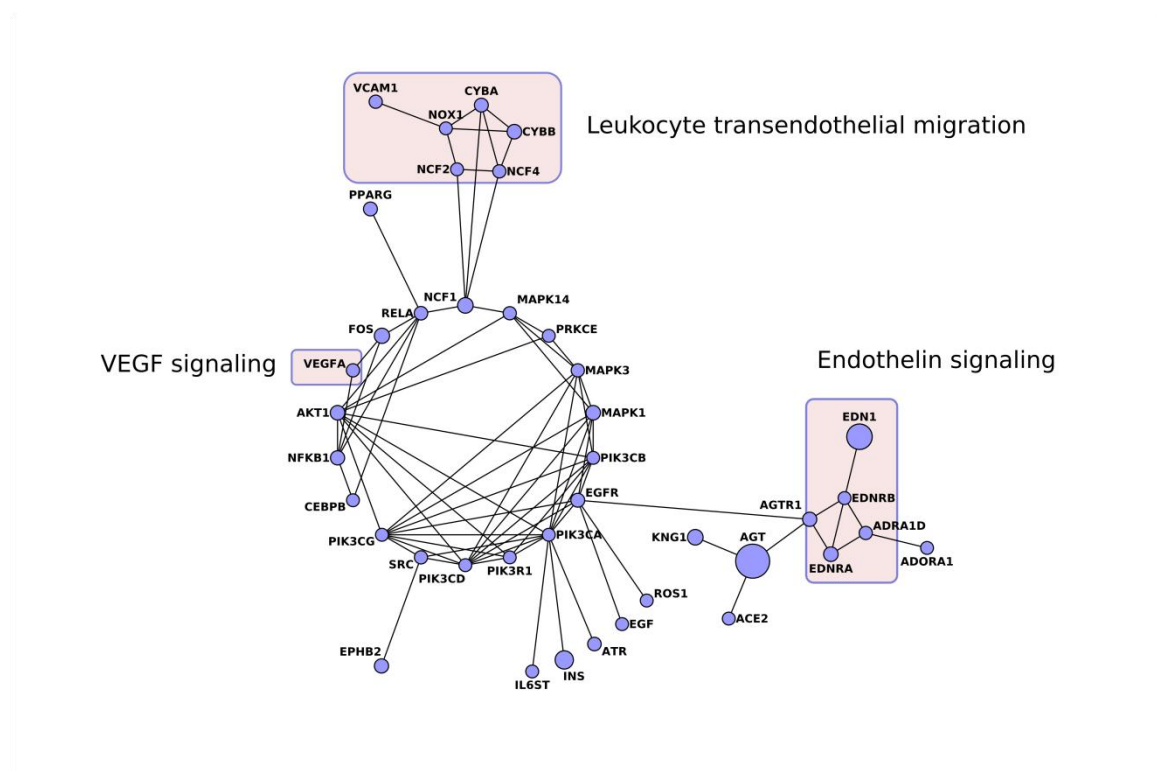


Figure 3: Protein interaction network of highly connected proteins associated with the cardiorenal syndrome. Nodes represent genes (gene symbols), edges indicate functional associations. Highlighted nodes represent proteins that are specific for either the VEGF signaling, the leukocyte transendothelial migration, or the endothelin signaling pathway.

Another highly connected cluster holds genes associated with leukocyte transendothelial migration. The process of leukocyte migration from blood into tissues is vital for inflammation, and it is known that inflammation is an important cardiorenal connector and a hallmark of kidney and heart diseases [5].

Identification of drug targets

116 out of the 280 proteins associated with the CRS were listed as drug target for at least one drug in DrugBank (see supplementary Table 1). The proteins with the most number of drugs were PTGS1, PTGS2, and NOS3 with 49, 43, and 41 drugs associated. The drug with the most drug targets in our list of 280 proteins was NADH.

Standard therapeutic regimes in the context of cardiovascular and kidney disease included aliskiren, irbesartan, or ramipril. Another drug candidate is nesiritide, a recombinant B-type natriuretic peptide that counter-regulates the RAS, as used in the treatment of acute decompensated heart failure (ADHF). However, on the basis of a prospective, randomized, double-blinded, placebo-controlled clinical trial Witteles et al. concluded that nesiritide therapy does not impact renal function in patients with ADHF and pre-existing renal dysfunction [47].

It is known that reducing blood pressure has beneficial effects on renal function and there is a multitude of antihypertensive agents acting on the RAS. Administration of angiotensin receptor antagonists in combination with angiotensin-converting enzyme inhibitors showed a significant reduction of urine albumin creatinine ratio in patients with hypertension and microalbuminuria and thus, a reduction of the risk for myocardial infarction [48].

Further potential targets for regulation of hemodynamics are members of the endothelin signaling pathway. Endothelin receptor antagonists are used in the treatment of a variety of cardiovascular conditions but less is known about the effects on combined kidney dysfunction. Ding et al. showed in animal models that chronic endothelin receptor blockade with endothelin receptor antagonists is beneficial in the treatment of progressive renal dysfunction and sodium retention associated with chronic heart failure [49]. Studies in humans are required to fully elucidate the effects and risks of endothelin receptor antagonist treatment in patients with CRS.

CONCLUSIONS

In this work we provide a comprehensive list of genes/proteins associated with the cardiorenal syndrome identified on the basis of a literature mining approach. On the basis of 825 articles identified in the context of CRS, 280 unique genes could be identified and were further characterized with respect to molecular function, biological processes, cellular pathways, subcellular location, tissue specific expression, as well as on the level of protein interaction networks.

The most frequently reported genes are involved in blood pressure regulating systems, particularly in the renin-angiotensin system (REN, AGT, ACE), as well as in the

antagonistic natriuretic peptide system (NPPA, NPPB). Enriched molecular functions include “receptor binding” and “receptor activity”. Of special note in this context are again players of the natriuretic peptide system, namely the two ligands NPPA and NPPB and its receptor NPR3. Tissue specific expression patterns of these molecules showed that NPPA and NPPB are mainly expressed in the heart, whereas their receptor NPR3 is highly expressed in kidney tissue, suggesting that this regulatory system is part of the crosstalk between the kidney and the cardiovascular system.

Therapy of the CRS is largely focused on natriuretic peptides or the renin-angiotensin system with a number of other molecular targets like the endothelin signaling pathway holding promise for future therapeutic strategies.

Altogether, the results of the present study strongly indicate the critical role of hemodynamic changes, blood pressure regulating hormone systems, and inflammatory processes in the formation of the CRS. Our analyses led to a comprehensive picture of molecular features involved in the functional interplay between the kidney and the cardiovascular system. One limitation of this automated literature mining approach is that we do not have experimental data on the expression levels of the reported molecules in the process of disease development. An obvious next step would therefore be to integrate the findings of this work with Omics datasets on kidney disease as well as vascular diseases. Such a combined approach has the potential to identify deregulated features for potentially identifying novel players for diagnostic or therapeutic approaches in the field of kidney and cardiovascular disease.

ACKNOWLEDGMENTS

This research was supported by Fresenius Medical Care Austria GmbH.

REFERENCES

1. Sarnak MJ, Levey AS, Schoolwerth AC, et al. Kidney disease as a risk factor for development of cardiovascular disease: a statement from the American Heart Association Councils on Kidney in Cardiovascular Disease, High Blood Pressure Research, Clinical Cardiology, and Epidemiology and Prevention. *Hypertension* **2003**; 42:1050-1065.
2. Ronco C, Haapio M, House AA, Anavekar N, Bellomo R. Cardiorenal syndrome. *J. Am. Coll. Cardiol* **2008**; 52:1527-1539.
3. Ronco C, McCullough PA, Anker SD, et al. Cardiorenal syndromes: an executive summary from the consensus conference of the Acute Dialysis Quality Initiative (ADQI). *Contrib Nephrol* **2010**; 165:54-67.
4. Schiffrin EL, Lipman ML, Mann JFE. Chronic kidney disease: effects on the cardiovascular system. *Circulation* **2007**; 116:85-97.
5. Mahapatra HS, Lalmalsawma R, Singh NP, Kumar M, Tiwari SC. Cardiorenal syndrome. *Iran J Kidney Dis* **2009**; 3:61-70.
6. Soni SS, Fahuan Y, Ronco C, Cruz DN. Cardiorenal syndrome: biomarkers linking kidney damage with heart failure. *Biomark Med* **2009**; 3:549-560.
7. Gutiérrez OM, Januzzi JL, Isakova T, et al. Fibroblast growth factor 23 and left ventricular hypertrophy in chronic kidney disease. *Circulation* **2009**; 119:2545-2552.
8. Chung AWY, Yang HHC, Kim JM, et al. Upregulation of matrix metalloproteinase-2 in the arterial vasculature contributes to stiffening and vasomotor dysfunction in patients with chronic kidney disease. *Circulation* **2009**; 120:792-801.
9. Perco P, Wilflingseder J, Bernthaler A, 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.
10. Jensen LJ, Saric J, Bork P. Literature mining for the biologist: from information retrieval to biological discovery. *Nat. Rev. Genet* **2006**; 7:119-129.
11. Krallinger M, Valencia A, Hirschman L. Linking genes to literature: text mining, information extraction, and retrieval applications for biology. *Genome Biol* **2008**; 9 Suppl 2:S8.
12. Yang Y, Adelstein SJ, Kassis AI. Target discovery from data mining approaches. *Drug Discov. Today* **2009**; 14:147-154.
13. Altman RB, Bergman CM, Blake J, et al. Text mining for biology--the way forward: opinions from leading scientists. *Genome Biol* **2008**; 9 Suppl 2:S7.
14. Agarwal P, Searls DB. Literature mining in support of drug discovery. *Brief. Bioinformatics* **2008**; 9:479-492.

15. Jani SD, Argraves GL, Barth JL, Argraves WS. GeneMesh: a web-based microarray analysis tool for relating differentially expressed genes to MeSH terms. *BMC Bioinformatics* **2010**; 11:166.
16. Lusis AJ, Weiss JN. Cardiovascular networks: systems-based approaches to cardiovascular disease. *Circulation* **2010**; 121:157-170.
17. Ashley EA, Ferrara R, King JY, et al. Network analysis of human in-stent restenosis. *Circulation* **2006**; 114:2644-2654.
18. Goh K-I, Cusick ME, Valle D, Childs B, Vidal M, Barabási A-L. The human disease network. *Proc. Natl. Acad. Sci. U.S.A* **2007**; 104:8685-8690.
19. Diez D, Wheelock AM, Goto S, et al. The use of network analyses for elucidating mechanisms in cardiovascular disease. *Mol Biosyst* **2010**; 6:289-304.
20. Weiss JN, Yang L, Qu Z. Systems biology approaches to metabolic and cardiovascular disorders: network perspectives of cardiovascular metabolism. *J. Lipid Res* **2006**; 47:2355-2366.
21. Barabási A-L, Oltvai ZN. Network biology: understanding the cell's functional organization. *Nat. Rev. Genet* **2004**; 5:101-113.
22. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabási AL. The large-scale organization of metabolic networks. *Nature* **2000**; 407:651-654.
23. Rizzuto R, Pinton P, Carrington W, et al. Close contacts with the endoplasmic reticulum as determinants of mitochondrial Ca²⁺ responses. *Science* **1998**; 280:1763-1766.
24. Vendelin M, Béraud N, Guerrero K, et al. Mitochondrial regular arrangement in muscle cells: a "crystal-like" pattern. *Am. J. Physiol., Cell Physiol* **2005**; 288:C757-767.
25. Hartwell LH, Hopfield JJ, Leibler S, Murray AW. From molecular to modular cell biology. *Nature* **1999**; 402:C47-52.
26. O'Rourke B, Ramza BM, Marban E. Oscillations of membrane current and excitability driven by metabolic oscillations in heart cells. *Science* **1994**; 265:962-966.
27. Aon MA, Cortassa S, O'Rourke B. Percolation and criticality in a mitochondrial network. *Proc. Natl. Acad. Sci. U.S.A* **2004**; 101:4447-4452.
28. Krallinger M, Valencia A. Text-mining and information-retrieval services for molecular biology. *Genome Biol* **2005**; 6:224.
29. Hu Y, Hines LM, Weng H, et al. Analysis of genomic and proteomic data using advanced literature mining. *J. Proteome Res* **2003**; 2:405-412.
30. Crim J, McDonald R, Pereira F. Automatically annotating documents with normalized gene lists. *BMC Bioinformatics* **2005**; 6 Suppl 1:S13.
31. McDonald R, Pereira F. Identifying gene and protein mentions in text using conditional random fields. *BMC Bioinformatics* **2005**; 6 Suppl 1:S6.

32. Diehn M, Sherlock G, Binkley G, et al. SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data. *Nucleic Acids Res* **2003**; 31:219-223.
33. Thomas PD, Campbell MJ, Kejariwal A, et al. PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res* **2003**; 13:2129-2141.
34. Thomas PD, Kejariwal A, Guo N, et al. Applications for protein sequence-function evolution data: mRNA/protein expression analysis and coding SNP scoring tools. *Nucleic Acids Res* **2006**; 34:W645-650.
35. Apweiler R, Bairoch A, Wu CH, et al. UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res* **2004**; 32:D115-119.
36. Horton P, Park K-J, Obayashi T, et al. WoLF PSORT: protein localization predictor. *Nucleic Acids Res* **2007**; 35:W585-587.
37. Schuler GD. Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *J. Mol. Med* **1997**; 75:694-698.
38. Bernthaler A, Mühlberger I, Fechete R, Perco P, Lukas A, Mayer B. A dependency graph approach for the analysis of differential gene expression profiles. *Mol Biosyst* **2009**; 5:1720-1731.
39. Wishart DS, Knox C, Guo AC, et al. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res* **2006**; 34:D668-672.
40. Wishart DS, Knox C, Guo AC, et al. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res* **2008**; 36:D901-906.
41. Judson WE, Helmer OM. Relationship of cardiorenal function to renin-aldosterone system in patients with valvular heart disease. *Circulation* **1971**; 44:245-253.
42. Ferrario CM. Addressing the theoretical and clinical advantages of combination therapy with inhibitors of the renin-angiotensin-aldosterone system: antihypertensive effects and benefits beyond BP control. *Life Sci* **2010**; 86:289-299.
43. McFarlane SI, Winer N, Sowers JR. Role of the natriuretic peptide system in cardiorenal protection. *Arch. Intern. Med* **2003**; 163:2696-2704.
44. Kang DH, Hughes J, Mazzali M, Schreiner GF, Johnson RJ. Impaired angiogenesis in the remnant kidney model: II. Vascular endothelial growth factor administration reduces renal fibrosis and stabilizes renal function. *J. Am. Soc. Nephrol* **2001**; 12:1448-1457.
45. Risau W. Mechanisms of angiogenesis. *Nature* **1997**; 386:671-674.
46. Rudnicki M, Perco P, Enrich J, 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. Witteles RM, Kao D, Christopherson D, et al. Impact of nesiritide on renal function in patients with acute decompensated heart failure and pre-existing renal dysfunction a randomized, double-blind, placebo-controlled clinical trial. *J. Am. Coll. Cardiol* **2007**; 50:1835-1840.

48. Menne J, Farsang C, Deák L, et al. Valsartan in combination with lisinopril versus the respective high dose monotherapies in hypertensive patients with microalbuminuria: the VALERIA trial. *J. Hypertens* **2008**; 26:1860-1867.
49. Ding S-S, Qiu C, Hess P, Xi J-F, Clozel J-P, Clozel M. Chronic endothelin receptor blockade prevents renal vasoconstriction and sodium retention in rats with chronic heart failure. *Cardiovasc. Res* **2002**; 53:963-970.

2.5.1 The Thesis Author's Contribution

Study design was predominantly the thesis author's responsibility. The thesis author further carried out the functional annotation and the analysis of the interaction network. Discussion and data interpretation were jointly done by all of the authors.

In detail, the following contributions are due to the thesis author's efforts:

- Design of the analysis workflow in consultation with other authors
- Selection of appropriate bioinformatics tools
- Functional annotation of genes derived from the literature mining approach, including biological process, molecular function and pathway enrichment analyses
- Extraction of tissue specific gene expression from the Unigene database
- Selection of relevant subgraphs resulting from the interaction analysis
- Discussion of genes frequently reported as associated with the cardiorenal syndrome, functional categories, relevant subgraphs and drug targets in collaboration with the other authors
- Visualization of subcellular location data and the protein interaction network
- Drafting the manuscript in cooperation with other authors

Molecular pathways and crosstalk characterizing the cardiorenal syndrome

Irmgard Mühlberger¹, Konrad Mönks¹, Raul Fechete¹, Gert Mayer², Rainer Oberbauer^{3,4}, Bernd Mayer^{1,5}, and Paul Perco^{1,4 *}

¹ emergentec biodevelopment GmbH, Gersthofer Strasse 29-31, 1180 Vienna, Austria

² Medical University of Innsbruck, Department of Internal Medicine IV, Anichstrasse 35, 6020 Innsbruck, Austria

³ Krankenhaus der Elisabethinen Linz, Fadingerstrasse 1, 4020 Linz, Austria

⁴ Medical University of Vienna, Department of Internal Medicine III, Waehringer Guertel 18-20, 1090 Vienna, Austria

⁵ Institute for Theoretical Chemistry, University of Vienna, Waehringer Strasse 17, 1090 Vienna, Austria

* Corresponding author:

Dr. Paul Perco
emergentec biodevelopment GmbH
Gersthofer Strasse 29-31
1180 Vienna, Austria
phone: +43-1-4034966
fax: +43-1-4034966-19
e-mail: paul.perco@emergentec.com

Submitted to: J Cell Mol Med. April 2011.

ABSTRACT

The risk of developing cardiovascular diseases (CVD) is dramatically increased in patients with chronic kidney diseases (CKD). Mechanisms leading to this cardiorenal syndrome (CRS) are multifactorial, and combined analyses of both failing organs may provide routes towards developing strategies for early risk assessment, prognosis, and consequently effective therapy.

In order to identify molecular mechanisms involved in the crosstalk between the diseased cardiovascular system and kidney, we analyzed tissue specific Omics profiles on atherosclerosis and diabetic nephropathy together with literature derived gene sets associated with cardiovascular and chronic kidney diseases. We focused on enriched molecular pathways and highlight molecular interactions found within as well as between affected pathways identified for the two organs.

Analysis on the level of molecular pathways points out the role of PPAR signaling, coagulation, inflammation, and focal adhesion pathways in formation and progression of the CRS. The proteins apolipoprotein A1 (APOA1) and albumin (ALB) turned out to be of particular importance in context of dyslipidemia, one of the major risk factors for the development of CVD.

In summary, our analyses highlight mechanisms associated with dyslipidemia, hemodynamic regulation, and inflammation on the interface between the cardiovascular and the renal system.

KEYWORDS: cardiorenal syndrome, chronic kidney disease, cardiovascular disease, literature mining, transcriptomics, pathways, protein interactions

INTRODUCTION

Patients suffering from chronic kidney disease are at high risk for developing cardiovascular complications. This fact already becomes evident for subjects with no or minor decrease in glomerular filtration rate (GFR) but showing protein excretion in urine with albuminuria being a strong predictor for cardiovascular complications. In end stage renal disease this relation becomes even more evident, with cardiovascular mortality being 10 to 30 times higher for patients on dialysis treatment compared to a matched general population with normal kidney function [1]. The clinical manifestations of cardiovascular disease in patients with kidney dysfunction are mainly atherosclerotic vascular disease and left ventricular hypertrophy [2]. A number of studies show that the prevalence of atherosclerosis is dramatically increased in dialysis patients and progressive over a range of reduced GFR [3-5]. Accelerated atherosclerosis can be frequently observed in diabetic nephropathy, being the leading cause of end-stage renal disease [6].

The pathophysiological state of combined kidney and cardiovascular dysfunction is described as cardiorenal syndrome (CRS), where the organ suffering in the first place can either be the cardiovascular system or the kidney. Further categorization depending on the origin of damage and the course of disease (either acute or chronic) has been established and discussed by Ronco and colleagues [7,8]. CRS 1 and 2 denote the acute or chronic cardio-renal syndrome respectively, whereas CRS 3 and 4 refer to reno-cardiac syndromes where the primary failing organ is the kidney. The fifth subtype characterizes cardio- and renal dysfunctions due to preceding systemic disorders such as sepsis or diabetes.

The mechanisms leading to all types of CRS are multifactorial and not restricted to changes of hemodynamic parameters like extracellular fluid volume, cardiac output, or arterial pressure only. Bongartz and colleagues outlined the four major cardiorenal connectors, namely increased activity of the renin-angiotensin system, oxidative stress, inflammation, and increased activity of the sympathetic nervous system [9]. Cardiac risk factors commonly associated with chronic kidney diseases, however, are complex and increase with age, the stage of kidney disease, and the level of proteinuria. Further factors include hypertension, diabetes, and dyslipidemia, and their appropriate treatment is certainly vital to reduce cardiovascular complications [10].

Early risk assessment and prognosis are key factors for effective and tailored treatment, particularly since management and therapy of severe cardiorenal syndrome is challenging. Therapeutic benefits of standard regimes are often achieved for one organ only or even worse, drugs in use for the treatment of cardiovascular diseases may go along with impairment of kidney function and vice versa. Further complications in treatment approaches leading to an increasing concern about novel strategies derive from the development of resistance to many standard therapies such as diuretics and inotropes [11]. So far, an effective therapy is lacking and further research, including the identification of biomarkers along with a better understanding of the underlying pathophysiological mechanisms to stratify CRS subtypes, is needed to develop selective therapeutic strategies.

In the last years, a significant number of genomics, transcriptomics, proteomics as well as metabolomics studies became available for characterizing altered kidney or cardiovascular function, but combined analyses of both failing organs on any omics level have been rare. One example is the gene expression analysis of aortic tissue from patients with or without chronic kidney disease scheduled for a coronary artery bypass graft, identifying differential expression of genes implicated in collagen fibrillogenesis and vascular smooth muscle cell migration [12].

An alternative approach for gaining a more global picture on disease mechanisms is the systematic extraction of information on genes and diseases as provided within the scientific literature. In particular, integrating results originating from different fields of research, such as e.g. cardiovascular disorders and kidney disease, represents a challenging task that can be facilitated by suitable literature mining methods. In this context, the most prevailing approach is based on concept co-occurrence as a measure for the relatedness of biomedical concepts (genes and associated diseases). We recently applied extensive literature mining on the CRS, identifying 280 unique genes/proteins discussed in this context. Analyzing these features on the level of protein interaction networks identified mediators of hemodynamic change as well as the endothelin and VEGF signaling pathway as centrally involved in the pathophysiology of CRS [13].

We in this work extend this literature mining approach by also including tissue specific Omics data sets. Transcriptomics profiles characterizing cardiovascular as well as

renal damage allow an integration of tissue-specific changes coupled with also systemic alterations covered by literature extraction methods. Specifically, we are interested in identifying pathways being jointly affected on the level of both organs for delineating molecular features potentially involved in molecular crosstalk of the cardiorenal syndrome.

METHODS

Data sets

Based on a catalogue of NCBI Medical Subject Headings (MeSH) specifying cardiovascular disease, renal disease, as well as the cardiorenal syndrome we extracted associated publications from Medline (database status as of April 2010). Subsequently, all genes associated with these publications were retrieved utilizing the gene-to-pubmed mapping file as provided by NCBI at <ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2pubmed.gz>, status as of April 2010. For identifying the significance of an association of a gene to a specific disease category we applied a Fisher's exact test using the number of associations of a given gene to a given disease category and the background distribution of gene-to-disease assignments as basis. Only genes showing a significant association with one of the diseases in focus ($p\text{-value} \leq 0.05$) were further considered.

Organ specific differentially expressed transcripts in chronic kidney disease and cardiovascular complications were extracted from two publicly available transcriptomics datasets. A first dataset published by Volger and colleagues provided profiles of human endothelial cells isolated from large arteries of patients with early and advanced atherosclerosis as compared to healthy controls. The list of differentially expressed genes of both, early and advanced atherosclerotic samples, as compared to control samples was retrieved from the supplementary material of the respective publication [14]. For generating a list of deregulated genes in chronic kidney disease we did make use of a publicly available dataset published by Schmid and colleagues on gene expression changes in human tubulointerstitial renal cells comparing patients with diabetic nephropathy and healthy controls [15]. This dataset was accessed through the Nephromine database (<http://www.nephromine.org>).

Next to the four described datasets, we included the literature-derived set of 280 proteins related to the cardiorenal syndrome as previously annotated and characterized in great detail by our group [13]. Molecular features identified as relevant via literature search and via Omics profile analysis were mapped to Entrez Gene IDs for allowing further joint analyses.

Protein interaction and pathway analysis

Protein-protein interactions (PPIs) between identified features were extracted from the IntAct database [16]. Feature sets were further mapped to extended KEGG pathways thus allowing an interpretation on a functional level [17]. Of the 214 pathways presently encoded in KEGG, 151 generic pathways were used, excluding all pathways specifically assigned to a disease phenotype. Pathways were extended as described in [18] to increase the coverage of genes assigned with these pathways, yielding a representation of 17,995 proteins.

For computing the enrichment of features assigned to cardiovascular or renal disease on the level of specific pathways a Fisher's exact test was used resting on the number of features assigned to a pathway and the number of features being identified as relevant for a given disease phenotype.

Molecular function and cellular component

Molecular features were annotated with respect to their molecular function and subcellular location according to the gene ontology database [19]. We specifically focused on the terms "receptor activity" and "extracellular space" for delineating the crosstalk between the two organs under study.

RESULTS

Identified molecular features

In total we identified 2,019 unique molecular features as being assigned to cardiovascular disorder, chronic kidney disease, and the cardiorenal syndrome. Table 1 provides an overview on the datasets and the numbers of molecular features identified.

dataset	Description	# features		
LIT-CVD	literature dataset based on cardiovascular MeSH terms	306	1386	2019
OMICS-CVD	deregulated transcripts derived from the atherosclerosis dataset	1096		
LIT-CKD	literature dataset based on chronic renal disease MeSH terms	183	540	
OMICS-CKD	deregulated transcripts derived from the diabetic nephropathy dataset	354		
CRS	literature dataset using the search term “cardiorenal”	280		

Table 1: Overview on datasets for cardiovascular disease (CVD) and chronic kidney disease (CKD) on the basis of literature extraction (LIT) and Omics data set analysis (OMICS).

The cardiovascular datasets held 306 (LIT-CVD) and 1,096 (OMICS-CVD) proteins respectively, with an overlap of 16 features (see figure 1A). From the 540 kidney disease specific features, 183 resulted from the literature mining approach (LIT-CKD) and 354 were part of the transcriptomics set (OMICS-CKD) (see figure 1B). Seven features were identified in both datasets, which is again a weak overlap but apparently not surprising, as the omics datasets are tissue specific, whereas literature mining also includes a systemic view on the different disease entities.

When merging literature and omics derived datasets, 101 features were found to be associated with kidney (CKD) and cardiovascular (CVD) diseases, and among these 30 features were also reported in the context of the cardiorenal syndrome (see figure 1C).

Furthermore, figure 1C gives information about organ-specific contributions. 43 genes assigned to cardiovascular complications and being part of the cardiorenal dataset appear not affected in kidney disease according to the CKD dataset. In turn, 13 members of the CRS dataset could be found in the kidney specific dataset but not in CVD specific profiles.

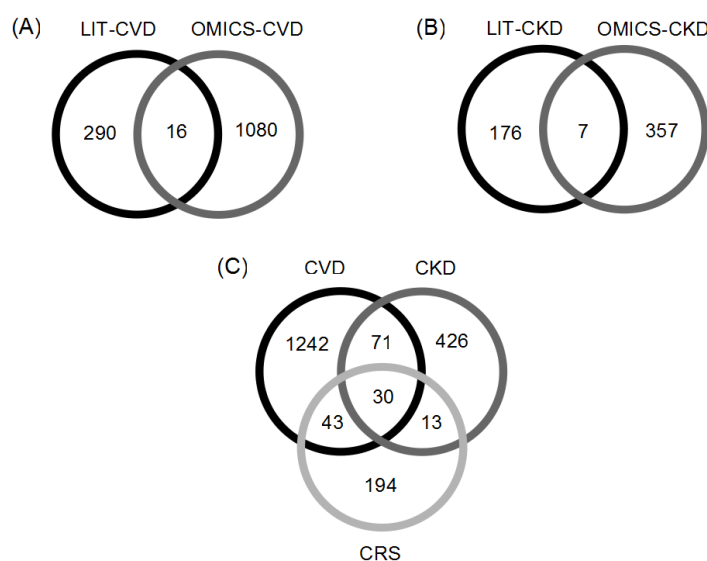


Figure 1: Venn diagrams showing the feature overlaps between (A) LIT-CVD and OMICS-CVD, (B) LIT-CKD and OMICS-CKD, and (C) CVD, CKD and CRS datasets.

Pathway analysis

Comparing the diseases on the level of affected pathways provided a more comprehensive picture than comparing individual features as such. We identified 29 enriched KEGG pathways for the literature and omics combined CVD and CKD features lists (see table 2). Joint pathways of both lists included the renin-angiotensin system, the complement and coagulation cascade, cytokine-cytokine receptor interactions, as well as the PPAR signaling, all of which were also significantly enriched by features of the CRS dataset. Three additional pathways were found to be coherently enriched within the CVD and CRS datasets, namely tyrosine metabolism, the adipocytokine signaling pathway, and vasopressin regulated water reabsorption. Focal adhesion was the only pathway jointly enriched by CKD and CRS specific genes.

pathway		# genes		
name	# genes total	CVD	CKD	CRS
renin-angiotensin system	50	8*	6*	5*
complement and coagulation cascades	180	38*	16*	9*
one carbon pool by folate	19	3*	1*	0
cytokine-cytokine receptor interaction	300	31*	16*	11*
PPAR signaling pathway	175	29*	10*	8*
glutathione metabolism	63	6*	1	2
purine metabolism	277	20*	5	6
glycosaminoglycan degradation	22	4*	1	0
tyrosine metabolism	34	2*	1	2*
circadian rhythm - mammal	36	3*	0	0
leukocyte transendothelial migration	59	7*	1	0
ubiquitin mediated proteolysis	662	39*	16	6
ABC transporters	97	11*	1	1
adipocytokine signaling pathway	126	15*	3	4*
hematopoietic cell lineage	65	4*	2	3
valine, leucine and isoleucine degradation	46	8*	1	0
gastric acid secretion	60	9*	0	3
vasopressin-regulated water reabsorption	152	14*	4	5*
lipoic acid metabolism	7	2*	0	0
mismatch repair	29	4*	2	0
arachidonic acid metabolism	58	11*	1	3
cardiac muscle contraction	35	10*	2	0
focal adhesion	149	13	9*	6*
pyruvate metabolism	37	1	3*	2
terpenoid backbone biosynthesis	28	1	1*	1
histidine metabolism	34	1	3*	0
O-Glycan biosynthesis	40	3	5*	0
chemokine signaling pathway	240	19	12*	5
ECM-receptor interaction	137	11	11*	3

Table 2: Enrichment of extended KEGG pathways for the CVD, CKD, and CRS datasets. Given is the pathway name, the total number of features assigned to the pathway following our extended pathway assignment, the number of features identified as relevant for CKD, CVD and CRS datasets, and significant enrichment for specific pathways and specific datasets (*).

Molecular crosstalk

Next to comparing CKD and CVD on the level of individual features as well as on the level of molecular pathways we analyzed evidence for specific protein interactions between members identified for CKD and CVD by mining the IntAct protein-protein interaction database. This procedure provided 284 protein-protein interactions identified for features associated with CVD or CKD. We specifically analyzed interactions between proteins that were assigned to enriched pathways in at least one of the two datasets, individually studying crosstalk i) between pathways enriched in both, CKD and CVD, ii) between pathways enriched in either CKD or CVD, and iii) within pathways enriched in both disease entities. Furthermore, information on the subcellular location of these proteins was added with particular focus on secreted proteins naturally being the most promising members when investigating the crosstalk between the two organs under study. The majority of interactors could actually be assigned to the GO category “extracellular space”. Proteins belonging to the GO molecular function category “receptor activity” were additionally marked.

For the crosstalk between pathways being enriched in both, CKD and CVD the complement and coagulation cascades and PPAR signaling were the two pathways showing the largest numbers of interconnected proteins. Key molecules of this crosstalk are apolipoprotein A1 (APOA1), being a known marker for cardiovascular disorders and apparently also affected in kidney diseases, as well as albumin (ALB). Albumin as a member of the CKD dataset interacts with the serpin peptidase inhibitor G1 (SERPING1), the coagulation factors II (F2) and VII (F7), and fibrinogen alpha (FGA), all members of the CVD dataset (see figure 2A), showing a further interaction to AHSG (Alpha2-HS glycoprotein) being assigned to the renin-angiotensin system.

Next, crosstalk between pathways exclusively enriched in either the cardiovascular or the kidney dataset was identified, holding proteins involved in ubiquitin mediated proteolysis, focal adhesion, and the complement and coagulation cascades, as outlined in figures 2B and 2C. Here a member of the ubiquitin mediated proteolysis (CVD enriched) links to focal adhesion (CKD enriched) as well as to a member situated in the complement and coagulation cascades (being enriched in both, CVD and CKD). PPAR signaling holds two members which on the one hand link to cardiac muscle contraction (CVD enriched) and on the other hand to focal adhesion (CKD enriched).

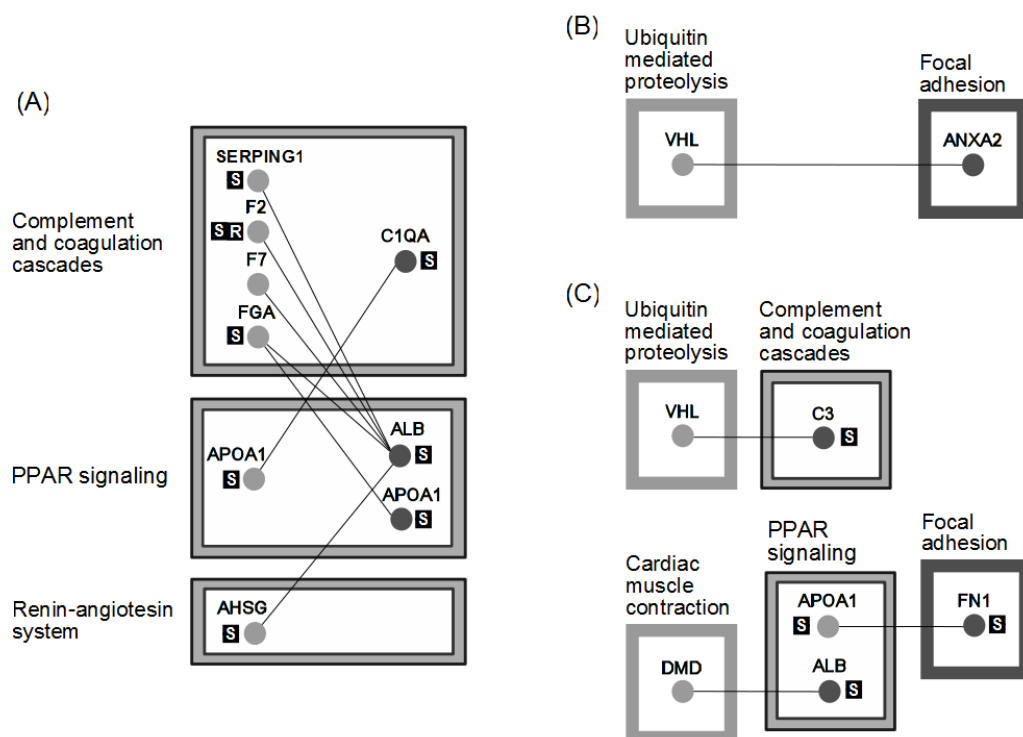


Figure 2: Inter-pathway links based on physical protein interactions for members from the CVD and CKD datasets. (A) Interactions of features found in pathways enriched by CVD as well as CKD features, (B) Interactions between features of pathways exclusively enriched by either the CVD (grey) or CKD (dark grey) dataset, and (C) links between pathways enriched by CVD as well as CKD features and pathways exclusively enriched by either the CVD or CKD dataset. 'S' and 'R' depicts secreted and receptor, respectively.

Looking at the intra-pathway protein interactions for pathways affected in both, CVD and CKD, a crosstalk between members of the CVD and CKD datasets was found for three pathways, namely PPAR signaling, the complement and coagulation cascade, as well as cytokine-cytokine receptor interactions (figure 3).

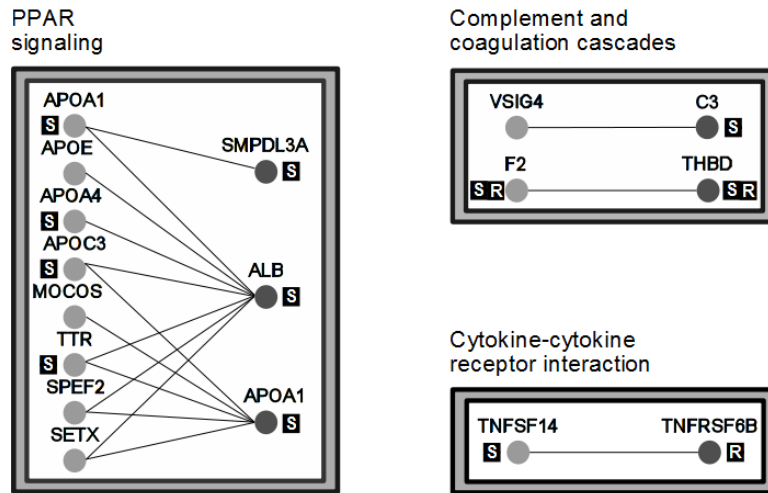


Figure 3: Intra-pathway links based on physical interactions found for members from the CVD and CKD datasets for pathways being affected in both, CVD and CKD. 'S' and 'R' depicts secreted and receptor, respectively.

Key players in the PPAR signaling pathway are found to be apolipoprotein A1 and albumin, showing in total fourteen links to other deregulated molecules. A receptor-ligand interaction in the complement and coagulation cascade was identified between the coagulation factor II and its binding partner thrombomodulin. Another receptor-ligand interaction between cardiovascular and kidney specific features was found in the context of cytokine-cytokine receptor interactions between the tumor necrosis factor receptor superfamily, member 6B (TNFRSF6B) and the tumor necrosis factor superfamily, member 14 (TNFSF14).

DISCUSSION

In the present paper we combined transcriptomics and scientific literature-derived datasets on chronic kidney disease and cardiovascular disease for characterizing the cardiorenal syndrome on a molecular pathway level. Following the scientific literature 280 features are directly reported in the context of CRS, and our integration of transcriptomics and literature data resulted in 1,386 features being linked with CVD, and 540 features linked with CKD. As frequently seen for such datasets on both, literature mining but in particular Omics datasets the overlap on the level of features is minor. This is, however, not surprising as e.g. the transcriptomics studies analyzed in this work are per definition tissue specific and analyze a specific clinical phenotype, which when combined with a more systems view on the disease as represented by scientific literature results in minor overlap. Going beyond the pure feature comparison to a more functional representation of disease pathology, as expressed by molecular pathways, frequently changes this picture.

We consequently investigated specifically whether dedicated pathways are found to be significantly populated by either kidney or cardiovascular disease specific features or both. Here we used a modified KEGG pathway set, where we on the one hand removed disease phenotype specific pathways as provided by KEGG, and on the other hand assigned proteins not embedded in KEGG according to a molecular relations approach. The resulting pathway map therefore focuses on key cellular processes further allowing an extended assignment of features given in our Omics- and literature-derived feature lists to such pathways.

Pathways affected in disease development in both organs were found as (i) the renin-angiotensin system, (ii) the complement and coagulation cascade, (iii) cytokine-cytokine receptor interactions, as well as (iv) the PPAR signaling pathway. These findings on the pathway level are in line with previously reported analysis results solely utilizing literature derived features associated specifically with the cardiorenal syndrome [13]. The connection between CRS and an increased activity of the blood pressure and water balance-regulating renin-angiotensin system was first reported in 1971 [20], and its consequences like renal hypoxia, vasoconstriction, intraglomerular hypertension, tubulointerstitial fibrosis, and proteinuria continue to be demonstrated in clinical practice. Members of the blood coagulation cascade are also heavily discussed

in the context of cardiovascular risk and renal diseases [21-24] also posing medical treatment options for both diseases [25]. A hypercoagulable state, often found in nephrotic CKD patients, is a major contributor to subsequent atherosclerosis and cardiovascular complications [26]. The recognition of cardiovascular, as well as renal protective properties of key players of the PPAR signaling pathway opens up further treatment options for the CRS. Next to the regulation of lipid concentrations in the blood, PPAR α / γ agonists exert anti-inflammatory and antioxidant actions [27]. They are widely used for the treatment of dyslipidemia as well as insulin resistance, and their beneficial effect on reducing arterial stiffness has been demonstrated in several clinical trials [28,29]. The positive effect of PPAR agonists on kidney function has so far been shown in animal models [30,31].

Altogether, jointly enriched pathways reflect several aspects of the pathophysiology of the CRS, including the dysregulation of hemodynamics, dyslipidemia, inflammation, and increased blood clotting, processes that are mainly addressed by current therapeutic strategies for the management of the CRS.

Next to the above discussed pathways that are affected in both organs we were interested in pathways exclusively enriched in either cardiovascular or renal datasets in order to draw conclusions on the organ specific contributions to the CRS. Pathways enriched in cardiovascular disease are found as associated with metabolism, hemodynamic regulation, vasopressin regulated water reabsorption, cardiac muscle contraction, as well as inflammation including the adipocytokine signaling pathway. Adipocytokine signaling is closely linked to the renin-angiotensin system and PPAR signaling pathways, as angiotensin II receptor blockers and PPAR α ligands improve the dysregulation of adipocytokine production, thereby reducing inflammation mediated changes [32,33]. Regulation of water reabsorption by vasopressin is achieved, among others, through fluid retention and activation of angiotensin II, thereby stimulating myocardial hypertrophy [34]. The beneficial effect of vasopressin antagonists on heart function without renal impairment has been reported recently [11].

Deregulated pathways in the renal system include cell adhesion, communication, as well as inflammation including the chemokine signaling pathway. Inflammation of renal tissue stimulates the expression of adhesion molecules in endothelial cells which, in turn, leads to the deposition of immune complexes and vascular stiffening in kidney disease [35].

A large number of the in total 27 protein-protein interactions found for features being associated with enriched pathways for cardiovascular and renal disease were detected within the PPAR signaling pathway as well as between members of the PPAR signaling pathways and members of the complement and coagulation cascade. Of major interest are those interactions where at least one of the interacting partners is secreted thus potentially mediating a direct cross-talk with the other organ. Major interactors of the PPAR signaling pathway are apolipoprotein A1 (APOA1) and albumin (ALB). Diabetic nephropathy is accompanied with dyslipidemia and, in contrast to most of the other apolipoproteins, decreased plasma levels of APOA1 [36]. Moreover, APOA1 values, particularly in relation to apolipoprotein B values, are used as estimates of cardiovascular risk [37]. APOA1 seems to be affected in both diseases and interacts with the complement component C1q (C1QA), which was found to be associated with chronic kidney disease. C1QA deficiency is associated with glomerular nephritis [38], but the relevance of the interactions of these proteins in the context of the CRS has not been evaluated yet. More is known about the interaction between APOA1 and the fibrinogen alpha chain (FGA). Studies in animal models outlined that the binding of apolipoprotein A to vessel walls via fibrinogen participates in the generation of atherosclerosis [39]. The direct interaction between APOA1 and fibronectin 1 (FN1), a member of the focal adhesion pathway, poses another interesting starting point for future research.

Albumin as the second major interactor derives from the CKD dataset and has a number of important functions. Hypoalbuminemia as a consequence of inflammation or loss in the urine in nephrotic kidney diseases has several consequences that can be associated with an increased cardiovascular risk, including a low osmotic pressure, an increased thromboembolic risk, and the accumulation of free fatty acids in the blood followed by an increased fibrinogen expression [40]. The strong connection between ALB and members of the coagulation cascade in the context of the CRS became evident by our findings. ALB and APOA1 also interact with a number of proteins associated with cardiovascular complications being also members of the PPAR signaling pathway. Therapeutically addressing the PPAR signaling appears promising for improving cardiovascular and chronic kidney disease.

Another interaction worth mentioning is found for the receptor TNFRSF6B and one of its ligands, TNFSF14, both members of the cytokine-cytokine receptor pathway. TNFSF14 associated signaling pathways are known to promote atherogenesis and are

suggested to be involved in chronic heart failure [41]. TNFRSF6B, which was found as associated to chronic kidney disease in our datasets is mainly reported in the context of cancer, and evaluating its role in kidney diseases and the relevance of TNFSF14 binding for formation or progression of the CRS requires further studies.

In a previous work on 280 literature derived proteins associated with the cardiorenal syndrome we identified hemodynamic changes, blood pressure regulating hormone systems, and inflammatory processes as central elements in the formation of the CRS, with a particular focus on the natriuretic peptide system, the renin-angiotensin system, and the endothelin signaling pathway [13]. In the present work we extended literature based datasets with transcriptomics profiles on kidney, as well as cardiovascular disease and could shed light on additional concepts like dyslipidemia and deregulated coagulation contributing to the CRS pathophysiology.

In summary, the consolidated analysis of tissue-specific changes together with systemic alterations covered by literature extraction methods for characterizing cardiovascular and kidney specific contributions to the CRS led to the identification of pathways relevant for disease formation and progression. Affected pathways are mainly associated with inflammation, cell adhesion, dyslipidemia, and hemodynamic regulation. First and foremost, PPAR signaling and the complement and coagulation cascade turned out to be significantly involved in disease mechanisms and thus, may be potential targets of therapeutic interventions. On a molecular level, our findings highlight the role of APOA1 and ALB as important molecules on the interface between the cardiovascular and the renal system.

ACKNOWLEDGMENTS

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

The work presented here was carried out in collaboration between all authors. IM, BM and PP designed the study with GM and RO further extending the study concept. IM, KM and RF contributed to data collection. IM performed the main analysis steps with contributions from RF. All authors contributed to data interpretation. IM, BM, and PP drafted the manuscript that was corrected, read and approved by all other authors.

The authors confirm that there are no conflicts of interest.

REFERENCES

1. Sarnak MJ, Levey AS, Schoolwerth AC, et al. Kidney disease as a risk factor for development of cardiovascular disease: a statement from the American Heart Association Councils on Kidney in Cardiovascular Disease, High Blood Pressure Research, Clinical Cardiology, and Epidemiology and Prevention. *Hypertension* **2003**; 42:1050-1065.
2. Berl T, Henrich W. Kidney-heart interactions: epidemiology, pathogenesis, and treatment. *Clin J Am Soc Nephrol* **2006**; 1:8-18.
3. Lindner A, Charra B, Sherrard DJ, Scribner BH. Accelerated atherosclerosis in prolonged maintenance hemodialysis. *N. Engl. J. Med* **1974**; 290:697-701.
4. Rostand SG, Kirk KA, Rutsky EA. The epidemiology of coronary artery disease in patients on maintenance hemodialysis: implications for management. *Contrib Nephrol* **1986**; 52:34-41.
5. Anavekar NS, McMurray JJV, Velazquez EJ, et al. Relation between renal dysfunction and cardiovascular outcomes after myocardial infarction. *N. Engl. J. Med* **2004**; 351:1285-1295.
6. Schiffrin EL, Lipman ML, Mann JFE. Chronic kidney disease: effects on the cardiovascular system. *Circulation* **2007**; 116:85-97.
7. Ronco C, Haapio M, House AA, Anavekar N, Bellomo R. Cardiorenal syndrome. *J. Am. Coll. Cardiol* **2008**; 52:1527-1539.
8. Ronco C, McCullough PA, Anker SD, et al. Cardiorenal syndromes: an executive summary from the consensus conference of the Acute Dialysis Quality Initiative (ADQI). *Contrib Nephrol* **2010**; 165:54-67.
9. Bongartz LG, Cramer MJ, Doevendans PA, Joles JA, Braam B. The severe cardiorenal syndrome: "Guyton revisited." *Eur. Heart J* **2005**; 26:11-17.
10. McCullough PA, Verrill TA. Cardiorenal interaction: appropriate treatment of cardiovascular risk factors to improve outcomes in chronic kidney disease. *Postgrad Med* **2010**; 122:25-34.
11. Koniari K, Nikolaou M, Paraskevaidis I, Parissis J. Therapeutic options for the management of the cardiorenal syndrome. *Int J Nephrol* **2010**; 2011:194910.
12. Fassot C, Briet M, Rostagno P, et al. Accelerated arterial stiffening and gene expression profile of the aorta in patients with coronary artery disease. *J. Hypertens* **2008**; 26:747-757.

13. Mühlberger I, Moenks K, Bernthaler A, et al. Integrative bioinformatics analysis of proteins associated with the cardiorenal syndrome. *Int J Nephrol* **2010**; 2011:809378.
14. Volger OL, Fledderus JO, Kisters N, et al. Distinctive expression of chemokines and transforming growth factor-beta signaling in human arterial endothelium during atherosclerosis. *Am. J. Pathol* **2007**; 171:326-337.
15. Schmid H, Boucherot A, Yasuda Y, et al. Modular activation of nuclear factor-kappaB transcriptional programs in human diabetic nephropathy. *Diabetes* **2006**; 55:2993-3003.
16. Kerrien S, Alam-Faruque Y, Aranda B, et al. IntAct--open source resource for molecular interaction data. *Nucleic Acids Res* **2007**; 35:D561-565.
17. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **2000**; 28:27-30.
18. Fechete R, Heinzl A, Perco P, et al. Mapping of molecular pathways, biomarkers and drug targets for diabetic nephropathy. *Proteomics Clin Appl* **2011**; [in press].
19. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet* **2000**; 25:25-29.
20. Judson WE, Helmer OM. Relationship of cardiorenal function to renin-aldosterone system in patients with valvular heart disease. *Circulation* **1971**; 44:245-253.
21. Kannel WB. Overview of hemostatic factors involved in atherosclerotic cardiovascular disease. *Lipids* **2005**; 40:1215-1220.
22. Kirmizis D, Tsiandoulas A, Pangalou M, et al. Validity of plasma fibrinogen, D-dimer, and the von Willebrand factor as markers of cardiovascular morbidity in patients on chronic hemodialysis. *Med. Sci. Monit* **2006**; 12:CR55-62.
23. Sjøland JA, Sidelmann JJ, Brabrand M, et al. Fibrin clot structure in patients with end-stage renal disease. *Thromb. Haemost* **2007**; 98:339-345.
24. Perco P, Mühlberger I, Mayer G, Oberbauer R, Lukas A, Mayer B. Linking transcriptomic and proteomic data on the level of protein interaction networks. *Electrophoresis* **2010**; 31:1780-1789.
25. Lynch AI, Boerwinkle E, Davis BR, et al. Antihypertensive pharmacogenetic effect of fibrinogen-beta variant -455G>A on cardiovascular disease, end-stage renal disease, and mortality: the GenHAT study. *Pharmacogenet. Genomics* **2009**; 19:415-421.
26. Morange PE, Bickel C, Nicaud V, et al. Haemostatic factors and the risk of cardiovascular death in patients with coronary artery disease: the AtheroGene study. *Arterioscler. Thromb. Vasc. Biol* **2006**; 26:2793-2799.

27. Schiffrin EL. More evidence of cardiorenal protective effects of peroxisome proliferator-activated receptor activation. *Hypertension* **2005**; 46:267-268.
28. Ryan KE, McCance DR, Powell L, McMahon R, Trimble ER. Fenofibrate and pioglitazone improve endothelial function and reduce arterial stiffness in obese glucose tolerant men. *Atherosclerosis* **2007**; 194:e123-130.
29. Werner C, Kamani CH, Gensch C, Böhm M, Laufs U. The peroxisome proliferator-activated receptor-gamma agonist pioglitazone increases number and function of endothelial progenitor cells in patients with coronary artery disease and normal glucose tolerance. *Diabetes* **2007**; 56:2609-2615.
30. Williams JM, Zhao X, Wang MH, Imig JD, Pollock DM. Peroxisome proliferator-activated receptor-alpha activation reduces salt-dependent hypertension during chronic endothelin B receptor blockade. *Hypertension* **2005**; 46:366-371.
31. Letavernier E, Perez J, Joye E, et al. Peroxisome proliferator-activated receptor beta/delta exerts a strong protection from ischemic acute renal failure. *J. Am. Soc. Nephrol* **2005**; 16:2395-2402.
32. Kurata A, Nishizawa H, Kihara S, et al. Blockade of Angiotensin II type-1 receptor reduces oxidative stress in adipose tissue and ameliorates adipocytokine dysregulation. *Kidney Int* **2006**; 70:1717-1724.
33. Toyoda T, Kamei Y, Kato H, et al. Effect of peroxisome proliferator-activated receptor-alpha ligands in the interaction between adipocytes and macrophages in obese adipose tissue. *Obesity (Silver Spring)* **2008**; 16:1199-1207.
34. Lee CR, Watkins ML, Patterson JH, et al. Vasopressin: a new target for the treatment of heart failure. *Am. Heart J* **2003**; 146:9-18.
35. Silverstein DM. Inflammation in chronic kidney disease: role in the progression of renal and cardiovascular disease. *Pediatr. Nephrol* **2009**; 24:1445-1452.
36. Attman PO, Knight-Gibson C, Tavella M, Samuelsson O, Alaupovic P. The compositional abnormalities of lipoproteins in diabetic renal failure. *Nephrol. Dial. Transplant* **1998**; 13:2833-2841.
37. Andrikoula M, McDowell IFW. The contribution of ApoB and ApoA1 measurements to cardiovascular risk assessment. *Diabetes Obes Metab* **2008**; 10:271-278.
38. Mitchell DA, Taylor PR, Cook HT, et al. Cutting edge: C1q protects against the development of glomerulonephritis independently of C3 activation. *J. Immunol* **1999**; 162:5676-5679.
39. Lou XJ, Boonmark NW, Horrigan FT, Degen JL, Lawn RM. Fibrinogen deficiency reduces vascular accumulation of apolipoprotein(a) and development of atherosclerosis in apolipoprotein(a) transgenic mice. *Proc. Natl. Acad. Sci. U.S.A* **1998**; 95:12591-12595.

40. Kim KJ, Yang WS, Kim SB, Lee SK, Park JS. Fibrinogen and fibrinolytic activity in CAPD patients with atherosclerosis and its correlation with serum albumin. *Perit Dial Int* **1997**; 17:157-161.
41. Dahl CP, Gullestad L, Fevang B, et al. Increased expression of LIGHT/TNFSF14 and its receptors in experimental and clinical heart failure. *Eur. J. Heart Fail* **2008**; 10:352-359.

2.6.1 The Thesis Author's Contribution

The thesis author designed the study in collaboration with other authors. Moreover, the author contributed to the data collection and performed functional analyses. Discussion and conclusions were due to the joint efforts of all of the authors.

In detail, the following contributions are due to the thesis author's efforts:

- Selection of transcriptomics datasets on diabetic nephropathy [15] and atherosclerosis [14] in consultation with other authors
- Development of the analysis workflow with the collaboration of other authors
- Selection of MeSH terms used for the extraction of genes from the literature
- Retrieval of transcriptomics datasets from the Nephromine database and respective publication
- Annotation of features with respect to their molecular function and subcellular location
- Identification and visualization of the direct feature overlap between the datasets
- Visualization of inter- and intra-pathway relationships based on protein interaction information from the IntAct database
- Discussion and interpretation of the results in communication with the other authors
- Lead in drafting the manuscript in cooperation with other authors

3. Discussion

3.1 Major Findings

The following sections summarize and discuss the major findings of the presented studies with a special focus on the results due to the thesis author's contributions.

3.1.1 Omics Workflows

Section 2.1 provides a detailed description of major omics data analysis steps covering data storage, retrieval, preprocessing, identification of differentially expressed features, functional annotation on the level of biological processes and molecular pathways, and interpretation of gene lists in the context of protein–protein interaction networks, as well as their exemplary application on a publicly available gene expression dataset on familial hypercholesterolemia.

The described workflows, including sequential, as well as integrated approaches were, among others, used for the analyses of different kidney diseases. A summary of the major findings resulting from the studies provided in sections 2.2 – 2.6 is given in the following chapters.

3.1.2 Acute Renal Failure/Transplantation

The studies presented in sections 2.2 and 2.3 cover the issue of ARF in the post-transplant situation.

Biomarkers derived from donor organs before engraftment may indicate the risk for developing ischemia reperfusion injury (IRI) and subsequent delayed graft function (DGF). Since graft failure is significantly more frequent after DGF compared to primary functioning grafts, the identification of subjects at risk before the event occurred is essential. The review “Biomarkers in Renal Transplantation Ischemia Reperfusion Injury” provides an overview on biomarker discovery and verification for the prediction of IRI and their utility for clinical use. An extensive literature search revealed 25

biomarker candidates presently discussed in the literature in the context of IRI and DGF, including Uromodulin (UMOD), hepatitis A virus cellular receptor 1 (HAVCR1, also known as kidney injury molecule 1 KIM1), and Cyclin-dependent kinase inhibitor (CDKN1A). Significantly enriched biological processes within the candidate list are, among others, angiogenesis and cell proliferation and differentiation. The latter, as well as processes associated with immunity and defense, were also found to be overpopulated by features lying on the shortest paths between the biomarker candidates in the protein interaction network omicsNet [39].

The same analysis procedure was repeated with a list of differentially expressed genes resulting from a transcriptomics study comparing live and deceased kidney donor organs. Immunity and defense processes were found to be significantly enriched by members of the original list of 90 differentially expressed genes, as well as by members of the subgraph representing the shortest paths between the differentially expressed genes according to omicsNet. Of particular interest are the subgraph members nuclear factor of kappa light polypeptide gene enhancer in B-cells 1 (NFKB1) and nuclear receptor subfamily 3, group C, member 1 (NR3C1), as these are targets for corticosteroids.

In order to test the hypothesis whether suppression of inflammation in the donor organ by steroids would ameliorate IRI and subsequently reduce the rate of DGF, a double-blinded, randomized, controlled trial was started. The outcomes of in total 455 transplant recipients receiving donor organs treated with either steroids or placebo showed no significant reduction in the incidence of DGF after steroid pretreatment [Kainz2010]. However, the functional enrichment analysis of differentially expressed genes in 20 steroid treated biopsies identified the up-regulation of inflammatory processes, limited transport capabilities, and a decreased metabolic activity of DGF organs compared to grafts with primary function. These results suggest a crucial role of hypoxia and it can be hypothesized that the activation of lipid and glucose metabolism may prevent the graft from developing ARF. Possible treatment strategies are the administration of peroxisome proliferator-activated receptor (PPAR) agonists or caspase inhibitors but further clinical trials are demanded to elucidate their beneficial effects on transplant outcome.

In summary, following the present data status inflammation events may be early stage indicators of IRI, triggering subsequent events along cell proliferation and apoptosis. However, a significant decrease of DGF could not be achieved by a steroid pretreatment of the donor organ. On a molecular level, inflammatory processes as well

as impaired transport and metabolic activities seem to distinguish delayed from primary functioning grafts.

3.1.3 Chronic Kidney Disease

The aim of the study presented in section 2.4 was the detection of coherences and differences between CKD specific kidney tissue transcriptomics and urine proteomics signatures.

Based on three transcriptomics and one proteomics dataset derived from urine, a number of analyses steps were performed on the level of direct feature overlap, biological processes, pathways, transcription factors, tissue expression, and interaction networks. The heterogeneity of the datasets became already evident when looking at the number of identified features, being 697 on part of the transcripts and 37 proteins. This large difference is certainly driven by the different sample matrix analyzed, as even in the presence of CKD only a limited number of proteins are released into the urine. Moreover, mRNA expression levels do not necessarily correlate with the respective protein abundance due to several reasons as regulatory mechanisms, post-translational modifications, pathophysiological conditions and so on. In view of these facts, the sparse overlap of only four features found when comparing transcriptomics and proteomics datasets is not surprising and leads to the assumption that an integrated analysis of omics profiles provides only moderate add-on information when solely aimed at identifying and subsequently correlating joint features.

However, the picture changes when going to the level of processes and pathways instead of comparing individual features as such. Of particular interest are the processes “cell structure and motility” and “immunity and defense”, as well as the pathways “ECM-receptor interaction”, “complement and coagulation cascades”, and “focal adhesion”, all of them found to be significantly enriched in both datasets. On the level of the omicsNet network, the role of hypercoagulability in disease formation could be further substantiated. Twelve members of the network spanned by 22 transcripts and 25 proteins showing strong inter-dependencies could be assigned to the GO term “coagulation”, including the serpin peptidase inhibitor C1 (SERPINC1) and the coagulation factors F2, F3, and F10. It is frequently reported that patients with CKD exhibit features of a hypercoagulable state which is also a main contributor to subsequent cardiovascular diseases.

3.1.4 Cardiorenal Syndrome

The high clinical relevance of the CRS due to the recognition of cardiovascular events as the leading cause of mortality in patients with chronic kidney diseases has driven a number of studies aiming at the identification of kidney-cardiovascular connectors. This rise in efforts is reflected by the growing number of publications associated with the keyword 'cardiorenal'.

The results of the analysis of 280 genes derived from 825 publications associated with the CRS are presented in section 2.5. The most frequently reported genes were found to be involved in blood pressure regulating systems, particularly in the renin-angiotensin system (renin REN, angiotensinogen AGT, angiotensin converting enzyme ACE), as well as in the antagonistic natriuretic peptide system NPS (natriuretic peptide A NPPA, natriuretic peptide B NPPB).

Enriched functional categories within the total set of 280 genes included "receptor binding" and "receptor activity". Following the assumption that CKD specific molecular features lead to alterations of the cardiovascular system and vice versa, the most probably scenario is the involvement of secretory features triggering receptor mediated downstream events in one of the affected organs. Actually, the classification of features in terms of subcellular location revealed "extracellular, including cell wall" as the most significantly enriched compartment. A specific example that perfectly matches the criteria for realizing the above mentioned scenario is the interplay between the natriuretic peptide receptor C (NPR3) and its ligands NPPA and NPPB. Tissue specific expression patterns of these molecules showed that NPPA and NPPB are mainly expressed in the heart, whereas their receptor NPR3 is highly expressed in kidney tissue.

The literature derived dataset covered the targets for most of the standard therapeutic regimes for the CRS to a great extend. Next to members of the RAS and NPS, features involved in the endothelin signaling pathway pose potential targets for drugs regulating hemodynamics.

In a next step, this literature mining approach was extended by also including tissue specific omics datasets (see section 2.6). Particularly, genes from publications that are tagged with CKD and CVD associated MeSH terms were extracted and combined with transcriptomics dataset on diabetic nephropathy and atherosclerosis.

Pathways identified as overpopulated by features specific for both diseases reflect several aspects of the pathophysiology of the CRS, including the dysregulation of

hemodynamics, dyslipidemia, inflammation, and increased blood clotting. Contributions on part of the cardiovascular system turned out to be mainly associated with hemodynamics and adipocytokine signaling, whereas the CKD specific signatures pointed towards the crucial role of impaired focal adhesion, chemokine signaling, and metabolic pathways in formation and progression of the CRS.

The investigation of inter- and intra-pathway relationships based on physical interaction information between CVD and CKD specific proteins showed an extensive organ crosstalk within the PPAR signaling pathway, as well as between members of the PPAR signaling pathway and the complement and coagulation cascade. Major interactors in this regard are, first and foremost, apolipoprotein A1 (APOA1) and albumin (ALB), as well as the complement component C1q (C1QA) and the fibrinogen alpha chain (FGA).

Therapeutically addressing the PPAR signaling system in case of dyslipidemia, insulin resistance or arterial stiffening is common, but its beneficial effect for specific treatment of the CRS needs further validation.

In summary, the literature mining approach has identified mediators of hemodynamic change, as well as the endothelin signaling pathway as centrally involved in the disease mechanisms of the CRS. Transcriptomics profiles characterizing cardiovascular as well as renal damage allowed an integration of tissue-specific changes coupled with also systemic alterations covered by literature extraction methods. This integrated approach could shed light on additional concepts like dyslipidemia and deregulated coagulation, contributing to CRS pathophysiology.

3.2 Outlook

Omics technologies have brought significant benefits in analysis and identification of molecular disease mechanisms, and opened up new opportunities for disease prediction, prevention, diagnosis, and treatment. Nevertheless, there are still a number of limitations in terms of technology, experimental design, statistical and functional analysis, validation, and clinical application that have to be seriously taken into account in order to obtain biologically meaningful results.

Until recently, the detection of splice variants by the usage of common microarray technologies was impossible. With regard to the high percentage of human genes that exhibit alternative splicing, the probability that a target sequence on the chip is not present in all forms of the respective transcript has to be considered. With the introduction of exon arrays which are designed to detect individual exons of a gene, possibilities for a quantitative assessment of transcripts comprehensively covering the human protein coding genome came up. One drawback of this new technology is certainly the availability of efficient tools for processing and analyzing the highly complex data. A further upcoming technology is tiling arrays. In fact, 60% of the transcriptional active regions in the human genome do not correspond to known exons. For example, non protein coding RNAs (ncRNAs), including structural RNAs (tRNAs, rRNAs, and snRNAs) and more recently discovered regulatory RNAs (e.g. microRNAs), fulfill a variety of important functions and were also found to be implicated in human diseases [67]. By offering the complete physical readout of a genome, tiling arrays can provide information outside of the boundaries of known protein coding genes.

Pivotal for any experiment is its reasoned design concerning collection and preparation of samples, as well as consideration of appropriate statistics.

First of all, omics studies in particular in translational clinical research should be done on precisely defined patient samples, and clinical parameters should be well matched in case/control designs. The same is true for cross-omics studies ideally to be done on the same patient status and, at best, on the same samples to assure maximum comparability. In this sense, collaborative efforts of research groups across different omics fields are demanded. One example is the large-scale integrating European project SysKid (Systems Biology towards Novel Chronic Kidney Disease Diagnosis and Treatment) which started in January 2010. SysKid integrates clinicians, statisticians, epidemiologists, molecular researchers across all omics fields, and bioinformaticians.

This interdisciplinary approach aims at understanding the pathophysiology of chronic kidney disease in order to provide tools both for identifying persons at risk for developing the disease, as well as for the development of novel therapy approaches (<http://www.syskid.eu>).

Another important issue to be considered in experimental design is the calculation of sample size. Omics experiments are often performed with a small number of samples due to their limited availability. However, only the inclusion of a sufficient number of independent samples provides the statistical power for the detection of true positive results.

The still improvable reproducibility of omics experiments, reflected by a usually weak overlap of results of individual studies on the very same study design, is frequently reported and can be attributed, next to variability in patient characteristics and weak statistical power due to small sample sizes, to several other factors. These include the use of different technical platforms, experimental variance including lack of uniform protocols, and the selection of different tools for data processing and statistical analysis [68].

Moreover, the validity of results can be influenced by sample heterogeneity that may lead to a high variability in gene expression measurements since expression can vary substantially among cell types. Thus, isolating specific cells of interest is an important step in sample preparation to prevent the detection of differences that may be unrelated to the biological question under study. Microdissection of e.g. kidney biopsy samples for specifically analyzing compartments of the kidney are an example.

Basically, the statistical analysis of omics data proceeds on the assumption that a maximal differential abundance of biological entities correlates with biological relevance. Particularly for regulatory elements like transcription factors, already minor changes in concentration may have significant impact on biological processes. Thus, such elements are less likely detected from a statistical perspective.

Concerning the functional analysis, the most limiting factor is certainly the incompleteness of existing annotation databases. This is particularly true for pathway data, as for example encoded by KEGG, that are far from being complete. An approach for computationally bypassing this issue was recently presented by Fechet et al. [69]. Functional enrichment analyses always have a certain annotation bias. If more data about a specific biological category is available, it is more likely to appear as significant than the others.

Molecular profiling via genomics, transcriptomics, proteomics or metabolomics has rapidly become the method of choice for biomarker discovery. However, despite promising results, only few novel biomarkers are yet used in clinical practice which is due to the long path from candidate discovery to qualification, verification, clinical validation, and finally commercialization. Not surprisingly, candidate biomarker discovery now commonly outruns the rate at which the candidates are being validated [70]. All the more so, the quality of candidate markers that are moved forward to the validation stage has to be ensured and can only be achieved by a profound data basis.

Even though several limitations of omics data generation, processing, analysis, and application exist, a large-scale approach on a Systems Biology level has the potential to provide insights in molecular processes contributing to kidney and cardiovascular disease formation and progression, as well as for the identification of biomarker candidates. Further advancements in sample work up, experimental technology, and analysis strategies, together with the establishment of collaborative networks and shared infrastructures for data and tools, will evolve our understanding of complex pathophysiological mechanisms, thereby assisting in the generation of hypothesis and leading to a more fundamental understanding of disease providing the basis for testing novel risk assessment, diagnosis, prognosis, and therapy options.

4. Appendix

References

1. Kitano H. Looking beyond the details: a rise in system-oriented approaches in genetics and molecular biology. *Curr. Genet* **2002**; 41:1-10.
2. Bruggeman FJ, Westerhoff HV. The nature of systems biology. *Trends Microbiol* **2007**; 15:45-50.
3. Rüegg C, Tissot J-D, Farmer P, Mariotti A. Omics meets hypothesis-driven research. Partnership for innovative discoveries in vascular biology and angiogenesis. *Thromb. Haemost* **2008**; 100:738-746.
4. Strange K. The end of “naive reductionism”: rise of systems biology or renaissance of physiology? *Am. J. Physiol., Cell Physiol* **2005**; 288:C968-974.
5. Barabási A-L, Oltvai ZN. Network biology: understanding the cell's functional organization. *Nat. Rev. Genet* **2004**; 5:101-113.
6. Hartwell LH, Hopfield JJ, Leibler S, Murray AW. From molecular to modular cell biology. *Nature* **1999**; 402:C47-52.
7. Almaas E. Biological impacts and context of network theory. *J. Exp. Biol* **2007**; 210:1548-1558.
8. Bernthaler A, Mühlberger I, Fechete R, Perco P, Lukas A, Mayer B. A dependency graph approach for the analysis of differential gene expression profiles. *Mol Biosyst* **2009**; 5:1720-1731.
9. Auffray C, Chen Z, Hood L. Systems medicine: the future of medical genomics and healthcare. *Genome Med* **2009**; 1:2.
10. Goh K-I, Cusick ME, Valle D, Childs B, Vidal M, Barabási A-L. The human disease network. *Proc. Natl. Acad. Sci. U.S.A* **2007**; 104:8685-8690.
11. Hocquette JF. Where are we in genomics? *J. Physiol. Pharmacol* **2005**; 56 Suppl 3:37-70.
12. Sanger F, Air GM, Barrell BG, et al. Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* **1977**; 265:687-695.
13. Kircher M, Kelso J. High-throughput DNA sequencing--concepts and limitations. *Bioessays* **2010**; 32:524-536.
14. Holley RW, Apgar J, Everett GA, et al. Structure of a ribonucleic acid. *Science* **1965**; 147:1462-1465.
15. Morozova O, Hirst M, Marra MA. Applications of new sequencing technologies for transcriptome analysis. *Annu Rev Genomics Hum Genet* **2009**; 10:135-151.

16. Freeman WM, Robertson DJ, Vrana KE. Fundamentals of DNA hybridization arrays for gene expression analysis. *BioTechniques* **2000**; 29:1042-1046, 1048-1055.
17. Brazma A, Hingamp P, Quackenbush J, et al. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet* **2001**; 29:365-371.
18. Desiere F, Deutsch EW, Nesvizhskii AI, et al. Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. *Genome Biol* **2005**; 6:R9.
19. The call of the human proteome. *Nat. Methods* **2010**; 7:661.
20. Cobon GS, Verrills N, Papakostopoulos P, Eastwood H, Linnane AW. The proteomics of ageing. *Biogerontology* **2002**; 3:133-136.
21. Hunter T, Andon N, Koller A, Yates J, Haynes P. The functional proteomics toolbox: methods and applications. *Journal of Chromatography B* **2002**; 782:165-181.
22. Aebersold R, Mann M. Mass spectrometry-based proteomics. *Nature* **2003**; 422:198-207.
23. Dhingra V, Gupta M, Andacht T, Fu Z. New frontiers in proteomics research: A perspective. *International Journal of Pharmaceutics* **2005**; 299:1-18.
24. Coon JJ, Zürlig P, Dakna M, et al. CE-MS analysis of the human urinary proteome for biomarker discovery and disease diagnostics. *Proteomics Clin Appl* **2008**; 2:964.
25. Wishart DS, Knox C, Guo AC, et al. HMDB: a knowledgebase for the human metabolome. *Nucleic Acids Res* **2009**; 37:D603-610.
26. Wishart DS, Tzur D, Knox C, et al. HMDB: the Human Metabolome Database. *Nucleic Acids Res* **2007**; 35:D521-526.
27. Oldiges M, Lütz S, Pflug S, Schroer K, Stein N, Wiendahl C. Metabolomics: current state and evolving methodologies and tools. *Appl. Microbiol. Biotechnol* **2007**; 76:495-511.
28. Krallinger M, Valencia A. Text-mining and information-retrieval services for molecular biology. *Genome Biol* **2005**; 6:224.
29. Krallinger M, Valencia A, Hirschman L. Linking genes to literature: text mining, information extraction, and retrieval applications for biology. *Genome Biol* **2008**; 9 Suppl 2:S8.
30. Chang JT, Schütze H, Altman RB. GAPSCORE: finding gene and protein names one word at a time. *Bioinformatics* **2004**; 20:216-225.
31. Tanabe L, Wilbur WJ. Tagging gene and protein names in biomedical text. *Bioinformatics* **2002**; 18:1124-1132.

32. Perco P, Rapberger R, Siehs C, et al. Transforming omics data into context: bioinformatics on genomics and proteomics raw data. *Electrophoresis* **2006**; 27:2659-2675.
33. Do JH, Choi D-K. Normalization of microarray data: single-labeled and dual-labeled arrays. *Mol. Cells* **2006**; 22:254-261.
34. Pan W. A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics* **2002**; 18:546-554.
35. Chen Y. Ratio-based decisions and the quantitative analysis of cDNA microarray images. *J. Biomed. Opt.* **1997**; 2:364.
36. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* **2009**; 37:1-13.
37. Stein L. Creating a bioinformatics nation. *Nature* **2002**; 417:119-120.
38. Weile J, Pocock M, Cockell SJ, et al. Customizable views on semantically integrated networks for systems biology. *Bioinformatics* **2011**; 27:1299-1306.
39. Bernthaler A, Mühlberger I, Fechete R, Perco P, Lukas A, Mayer B. A dependency graph approach for the analysis of differential gene expression profiles. *Mol Biosyst* **2009**; 5:1720-1731.
40. von Mering C, Jensen LJ, Kuhn M, et al. STRING 7--recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res* **2007**; 35:D358-362.
41. Szklarczyk D, Franceschini A, Kuhn M, et al. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res* **2011**; 39:D561-D568.
42. Rennke HG, Denker BM, Rose BD. *Renal pathophysiology: the essentials*. Lippincott Williams & Wilkins, 2007.
43. Nissenson AR. Acute renal failure: definition and pathogenesis. *Kidney Int. Suppl* **1998**; 66:S7-10.
44. Brenner BM. *Brenner and Rector's the Kidney*. Philadelphia: Saunders W B Co, 2007.
45. Bellomo R, Ronco C, Kellum JA, Mehta RL, Palevsky P. Acute renal failure - definition, outcome measures, animal models, fluid therapy and information technology needs: the Second International Consensus Conference of the Acute Dialysis Quality Initiative (ADQI) Group. *Crit Care* **2004**; 8:R204-212.
46. Needham E. Management of acute renal failure. *Am Fam Physician* **2005**; 72:1739-1746.
47. Ojo AO, Wolfe RA, Held PJ, Port FK, Schmodder RL. Delayed graft function: risk factors and implications for renal allograft survival. *Transplantation* **1997**; 63:968-974.

48. Kainz A, Mitterbauer C, Hauser P, et al. Alterations in gene expression in cadaveric vs. live donor kidneys suggest impaired tubular counterbalance of oxidative stress at implantation. *Am. J. Transplant* **2004**; 4:1595-1604.
49. Hauser P, Schwarz C, Mitterbauer C, et al. Genome-wide gene-expression patterns of donor kidney biopsies distinguish primary allograft function. *Lab. Invest* **2004**; 84:353-361.
50. Dennen P, Parikh CR. Biomarkers of acute kidney injury: can we replace serum creatinine? *Clin. Nephrol* **2007**; 68:269-278.
51. Lisowska-Myjak B. Serum and urinary biomarkers of acute kidney injury. *Blood Purif* **2010**; 29:357-365.
52. Perco P, Pleban C, Kainz A, et al. Protein biomarkers associated with acute renal failure and chronic kidney disease. *Eur. J. Clin. Invest* **2006**; 36:753-763.
53. Soni SS, Ronco C, Katz N, Cruz DN. Early diagnosis of acute kidney injury: the promise of novel biomarkers. *Blood Purif* **2009**; 28:165-174.
54. Levey AS, Coresh J, Balk E, et al. National Kidney Foundation practice guidelines for chronic kidney disease: evaluation, classification, and stratification. *Ann. Intern. Med* **2003**; 139:137-147.
55. Snyder S, Pendergraph B. Detection and evaluation of chronic kidney disease. *Am Fam Physician* **2005**; 72:1723-1732.
56. Theilig F. Spread of glomerular to tubulointerstitial disease with a focus on proteinuria. *Ann. Anat* **2010**; 192:125-132.
57. Abbate M, Zoja C, Remuzzi G. How Does Proteinuria Cause Progressive Renal Damage? *Journal of the American Society of Nephrology* **2006**; 17:2974-2984.
58. Liu B-C, Lü L-L. Novel biomarkers for progression of chronic kidney disease. *Chin. Med. J* **2010**; 123:1789-1792.
59. Sarnak MJ, Levey AS, Schoolwerth AC, et al. Kidney disease as a risk factor for development of cardiovascular disease: a statement from the American Heart Association Councils on Kidney in Cardiovascular Disease, High Blood Pressure Research, Clinical Cardiology, and Epidemiology and Prevention. *Hypertension* **2003**; 42:1050-1065.
60. Ronco C, Haapio M, House AA, Anavekar N, Bellomo R. Cardiorenal syndrome. *J. Am. Coll. Cardiol* **2008**; 52:1527-1539.
61. Ronco C, McCullough PA, Anker SD, et al. Cardiorenal syndromes: an executive summary from the consensus conference of the Acute Dialysis Quality Initiative (ADQI). *Contrib Nephrol* **2010**; 165:54-67.
62. Breidthardt T, Mebazaa A, Mueller CE. Predicting progression in nondiabetic kidney disease: the importance of cardiorenal interactions. *Kidney Int* **2009**; 75:253-255.
63. García-López E, Carrero JJ, Suliman ME, Lindholm B, Stenvinkel P. Risk factors for cardiovascular disease in patients undergoing peritoneal dialysis. *Perit Dial Int* **2007**; 27 Suppl 2:S205-209.

64. Liu Y, Berthier-Schaad Y, Fallin MD, et al. IL-6 haplotypes, inflammation, and risk for cardiovascular disease in a multiethnic dialysis cohort. *J. Am. Soc. Nephrol* **2006**; 17:863-870.
65. McIntyre CW. Effects of hemodialysis on cardiac function. *Kidney Int* **2009**; 76:371-375.
66. McFarlane SI, Winer N, Sowers JR. Role of the natriuretic peptide system in cardiorenal protection. *Arch. Intern. Med* **2003**; 163:2696-2704.
67. Zhang Z, Pang AWC, Gerstein M. Comparative analysis of genome tiling array data reveals many novel primate-specific functional RNAs in human. *BMC Evol. Biol* **2007**; 7 Suppl 1:S14.
68. Weintraub LA, Sarwal MM. Microarrays: a monitoring tool for transplant patients? *Transpl. Int* **2006**; 19:775-788.
69. Fechete R, Heinzel A, Perco P, et al. Mapping of molecular pathways, biomarkers and drug targets for diabetic nephropathy. *Proteomics Clin Appl* **2011**; [in press].
70. Lin D, Hollander Z, Meredith A, McManus BM. Searching for “omic” biomarkers. *Can J Cardiol* **2009**; 25 Suppl A:9A-14A.

Abstract

Kidney diseases represent a significant health burden with a number of currently unmet clinical needs in both, diagnosis/prognosis as well as therapy. Epidemiological studies show that about 10% of the general population suffers from early stages of reduced kidney function, contributing to bone metabolism disorders and cardiovascular complications. In the realm of 'omics' approaches a significant number studies have been driven by various groups for characterizing altered kidney function, and singular analyses of such profiles have provided insight into processes of inflammation and hemodynamic regulation as central elements for contributing to the pathophysiology of kidney diseases. However, an integrated analysis of kidney diseases in the spirit of Systems Biology is still in its infancy.

Following the evident clinical needs and methodological shortcomings on analyzing and understanding diseases of the kidney, this thesis addresses sequential analysis procedures from data processing to functional analyses of large scale transcriptomics data, as well as integrated workflows for handling and cross-linking multi-level omics data primarily in the context of protein interaction networks. Conceptual development in this area was then tested by using available omics data on various forms of kidney disease.

The combined analysis of literature- and transcriptomics-based genes shed light on molecular links between the cardiovascular system and chronic diseased kidneys and thus, allowed the identification of potential novel therapeutic targets addressing the cardiorenal syndrome. Further analysis concerning end-stage renal diseases, particularly the post-transplant situation, revealed a set of biomarker candidates that promise early risk assessment of a delayed graft function, including VEGF and CDKN1A. On a molecular level, inflammation events turned out to be early-stage indicators for kidney function. However, results of a randomized control trial showed no reduction of the rate of delayed graft function after steroid pretreatment of donor organs.

An integrated analysis workflow following a Systems Biology approach, as exemplified in this thesis, has the potential for identifying molecular processes contributing to disease formation and progression, biomarker candidates for diagnosis and risk assessment, as well as for generating hypothesis leading to a more fundamental

understanding of disease mechanisms providing the basis for testing novel therapy options.

Zusammenfassung

Nierenerkrankungen stellen eine erhebliche gesundheitliche Belastung dar, und Verbesserung in Diagnose, Prognose und Therapie sind zentrale Elemente. Epidemiologische Studien zeigen, dass etwa 10% der Gesamtbevölkerung an den ersten Zeichen einer eingeschränkten Nierenfunktion leidet, und dies wiederum erhöht das Risiko für Knochenstoffwechselerkrankungen und Herz-Kreislauf-Komplikationen. In den letzten Jahren gab es eine Vielzahl an Studien die das Ziel hatten, mit Hilfe von Omics-Technologien die Veränderung der Nierenfunktion zu charakterisieren. Die Ergebnisse aus den Analysen von einzelnen Omics-Profilen lassen darauf schließen, dass ein maßgeblicher Beitrag zur Pathophysiologie der Nierenerkrankung von entzündlichen Prozessen und hämodynamischer Fehlregulation stammt. Integrative Analysen im Sinne der Systembiologie stecken allerdings noch in den Anfängen.

Diese vorliegende Arbeit umfasst sequentielle und integrative Analyseverfahren von Omics-Daten zu verschiedenen Arten der Nierenerkrankung um sowohl methodologisch wie auch klinisch zu den gegebenen Fragestellungen beizutragen. Diese beinhalten das Prozessieren und die funktionale Analyse von Genexpressionsdaten, bis hin zur Handhabung und Verknüpfung von heterogenen Omics-Daten auf der Basis von Proteininteraktionsnetzwerken.

Ergebnisse aus der Analyse von relevanten Genen aus Literatur und aus Genexpressionsdaten zeigten molekulare Verbindungen zwischen dem Herz-Kreislauf-System und der chronischen Nierenerkrankung auf („Kardioresnales Syndrom“), die des Weiteren auch zur Identifikation von potentiellen neuen Angriffspunkten für therapeutische Maßnahmen führten. Durch weitere Analysen zu Nierenerkrankungen im Endstadium, fokussiert auf die Post-Transplant-Situation, konnten eine Reihe von Biomarker Kandidaten abgeleitet werden, die eine frühe Risikoabschätzung hinsichtlich verzögerter Transplantatfunktion versprechen, darunter VEGF und CDKN1A. Grundsätzlich zeigen die Analysen, dass Entzündungsprozesse auf molekularer Ebene sehr frühe Indikatoren hinsichtlich einer Einschränkung der Nierenfunktion darstellen. Eine randomisierte, kontrollierte Studie konnte allerdings keine Abnahme der Zahl an Transplantaten mit verzögerter Funktion nach Vorbehandlung des Spenderorgans mit Steroiden bestätigen.

Integrative Analyseabläufe in einem systembiologischen Ansatz, so wie in dieser Arbeit beschrieben, haben das Potential molekulare Prozesse zu identifizieren die an Krankheitsentstehung und Progression beteiligt sind, Biomarkerkandidaten für Diagnose und Risikoabschätzung hervorzubringen, und Hypothesen zu generieren, die zu einem besseren Verständnis der Krankheitsmechanismen führen und somit die Basis für das Testen von neuen Therapieoptionen darstellen.

Curriculum Vitae

PERSONAL DATA

<i>Name</i>	Mag. Irmgard Mühlberger
<i>Email</i>	irmgard.muehlberger@gmx.at
<i>Date of birth</i>	10.09.1981
<i>Nationality</i>	Austria

EDUCATION

<i>Since July 2008</i>	Dissertation at the Institute of Theoretical Chemistry, University of Vienna
<i>Jun 2008</i>	Diploma in molecular biology
<i>Sep 2007-Jun 2008</i>	Diploma work at the Institute of Theoretical Chemistry, University of Vienna
<i>Mar 2006-Jun 2007</i>	Study of informatics at the Technical University of Vienna
<i>Mar-Apr 2005</i>	Internship at the Institute of Microbiology and Genetics, University of Vienna
<i>Feb 2005</i>	Internship at the Institute of Theoretical Chemistry, University of Vienna
<i>Since Oct 2001</i>	Study of molecular biology at the University of Vienna
<i>Jun 2000</i>	A-level
<i>Sep 1991-Jun 2000</i>	Secondary School (GRG15, Auf der Schmelz)

PROFESSIONAL EXPERIENCE

<i>Sep 2007-Dez 2008</i>	Emergentec biodevelopment GmbH, Research and Development
<i>Jan 2009-Oct 2009</i>	Medical University of Innsbruck, Research and Development
<i>Nov 2009-April 2011</i>	Emergentec biodevelopment GmbH, Research and Development

POSTER PRESENTATIONS

June 2008

6th International Conference on Pathways, Networks and Systems
Medicine
Chania, Crete, Greece, June 16-21, 2008

A dependency graph approach for analyzing differential gene expression data of B-cell lymphomas.

I. Mühlberger, P. Perco, A. Bernthaler, R. Fechete, A. Lukas, and B. Mayer

June 2009

ISMB/ECCB 2009
Stockholm, Sweden, June 27 – July 2, 2009

Omics profile integration for characterizing kidney diseases.

I. Mühlberger, P. Perco, A. Bernthaler, R. Fechete, K. Mönks, R. Oberbauer, G. Mayer, A. Lukas, and B. Mayer

PUBLICATIONS

Hauser PV, Perco P, Mühlberger I, Pippin J, Blonski M, Mayer B, Alpers CE, Oberbauer R, Shankland SJ. **Microarray and bioinformatics analysis of gene expression in experimental membranous nephropathy.** *Nephron Exp Nephrol.* 2009;112(2):e43-58.

Bernthaler A, Mühlberger I, Fechete R, Perco P, Lukas A, Mayer B. **A dependency graph approach for the analysis of differential gene expression profiles.** *Mol Biosyst.* 2009;5(12):1720-31.

Mühlberger I, Perco P, Fechete R, Mayer B, Oberbauer R. **Biomarkers in renal transplantation ischemia reperfusion injury.** *Transplantation.* 2009;88(3 Suppl):S14-19.

Perco P, Mühlberger I, Mayer G, Oberbauer R, Lukas A, Mayer B. **Linking transcriptomic and proteomic data on the level of protein interaction networks.** *Electrophoresis.* 2010;31(11):1780-9.

Wilflingseder J, Kainz A, Mühlberger I, Perco P, Langer R, Kristo I, Mayer B, Oberbauer R. **Impaired metabolism in donor kidney grafts after steroid pretreatment.** *Transpl Int.* 2010;23(8):796-804.

Mühlberger I, Moenks K, Bernthaler A, Jandrasits C, Mayer B, Mayer G, Oberbauer R, Perco P. **Integrative bioinformatics analysis of proteins associated with the cardiorenal syndrome.** *Int J Nephrol.* 2010;2011:809378.

Mühlberger I, Wilflingseder J, Bernthaler A, Fechete R, Lukas A, Perco P. **Computational analysis workflows for Omics data interpretation.** *Bioinformatics for Omics Data: Methods and Protocols*, B. Mayer ed., *Methods in Molecular Biology* Vol 719, Humana Press NY, 2011.

Mühlberger I, Mönks K, Fechete R, Mayer G, Oberbauer R, Mayer B, Perco P. **Molecular pathways and crosstalk characterizing the cardiorenal syndrome.** Submitted to *J Cell Mol Med*, 2011.