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

Paul Perco[1], Irmgard Mühlberger[1], Gert Mayer[2],
Rainer Oberbauer[3], Arno Lukas[1], Bernd Mayer[1,*]

[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

Keywords: data integration, proteomics, transcriptomics, networks, renal disease

Word count: 4897, 4 tables, 5 figures

* Corresponding author:
Bernd Mayer
Tel/Fax: +43-1-4034966
e-mail: bernd.mayer@emergentec.com

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 an analysis strategy for integrating omics profiles on the level of protein interaction networks, exemplified on 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 omics data sources on a given phenotype under analysis on directed but also undirected protein interaction networks serves in joint functional interpretation of transcriptomics and proteomics data sets.

1. 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 comparably large 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 are applied for protein identification and quantization. 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 in the form of controlled vocabularies for enabling standardized retrieval and analysis.

However, most analysis performed is ‘within a domain’, i.e. transcriptomics and proteomics analysis follows established analysis pipelines basically aimed at deriving abundance profiles 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, is unfortunately less frequently done. However, utilizing resources as the Gene Expression Omnibus and PRIDE allows extracting both data levels for a number of cellular conditions, in principal allowing a joint analysis of both profiles.

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 concentrations as derived by 2D-PAGE they concluded that a simple deduction of protein concentrations from mRNA transcript analysis was 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, and they reported that 73% of the variance in yeast protein abundance is explained by mRNA concentration [7]. In a recent study by Shankavaram and colleagues on the large NCI-60 cancer cell panel around two thirds (65%) of the genes in the dataset showed statistically significant transcript-protein correlations [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, including cell adhesion and protein folding complexes.

Reasons for a poor correlation between mRNA and protein abundance may be classified into two major groups (Table 1). The first group (intrinsic mechanisms) consists of regulatory mechanisms in the course of gene expression, as well as protein modifications, both being responsible for a poor correlation although the detected proteins are a consequence of the measured mRNA. These mechanisms include transcriptional and translational regulations, as well as post-translational modifications and variable protein half-life times *in-vivo*.

The second group deals with downstream events rather than the process of gene expression itself (extrinsic mechanisms). Disease mechanisms can lead to high amount of proteins in specific tissues although the protein synthesis rate is not affected. 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.

Another factor to be included is the identification of secreted proteins in certain tissues and body fluids whose transcripts are prevalence in a different tissue [10]. 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 correlations 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. 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 these profiles for circumventing intrinsic and extrinsic causes for failure of direct correlation. 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 m^2 with considerable variation between and even within individuals. Below 60 ml/min/1.73 m^2 the prevalence of complications of CKD increases, the risk of cardiovascular events is elevated even at earlier stages. The most severe form of CKD is end-stage renal disease with the treatment options of dialysis or transplantation. Transcriptomics as well as proteomics methodologies have significantly contributed toward unraveling molecular mechanisms leading to CKD [17-19], and linking available omics profiles promises a further understanding of this disease.

2. MATERIALS AND METHODS

Data sets

We used three publicly available microarray studies on chronic kidney disease for the generation of the list of deregulated features on the mRNA level. Two studies focused on differences in mRNA expression in diabetic nephropathy using Affymetrix Gene Chips. Schmid and colleagues compared mRNA levels in the tubulointerstitial compartment of thirteen diseased patients and seven healthy control subjects [20]. The second dataset was published by Baelde and colleagues and holds transcripts differentially expressed between cells of glomeruli from two diseased and two morphologically normal kidneys [21]. The third study by Rudnicki and colleagues 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 [17]. Spotted cDNA microarrays from the Stanford Functional Genomics Facility were used in the analysis.

The proteomics dataset was extracted from the Human Urinary Proteome Database (HUPDB), which 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, IgA nephropathy, membranous glomerulonephritis, focal segmental glomerulosclerosis, and minimal change disease, all forms of chronic kidney disease. Only those proteins which were present in more than 30% of the urine samples in one disease group were further investigated.

Analysis procedures

Differentially regulated transcripts and proteins were mapped to their respective NCBI Gene Symbols in order to make the transcriptomics and the proteomics lists comparable. In a first step those features present in both lists were identified. In successive analyses the lists were interpreted on the level of functional annotation, molecular pathways and protein dependency networks.

Functional annotation

Enriched biological processes for both the transcriptomics and the proteomics list were identified using the PANTHER (Protein Analysis THrough Evolutionary Relationships) Classification System [23]. PANTHER is an ontology where 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 [24, 25]. 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 (splice variants are implicitly encoded). Edges between nodes represent pairwise dependencies which were calculated by integrating similarity and functional dependency measures. These measures 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. We used omicsNET in order to identify strong dependencies between transcripts and proteins thus showing edge weights of two or above (where the edge weights scale in-between -1.8 and 5.4, where a value of 5.4 represents maximum dependency of a given pair). We further focused on members of a specific biological process and retrieved omicsNET subgraphs holding only transcripts and proteins involved in the blood coagulation cascade showing dependencies above certain thresholds. We used three edge weight cutoff values of 1, 1.5 and 2 respectively in order to generate the networks.

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 datasest. MAPPER is a database holding information on putative transcription factor binding sites in the regulatory regions of genes in various species [26].

Kidney tissue expression

Data on immunohistochemical staining in renal tissue were retrieved from the Human Protein Atlas for the features in the proteomics dataset. The Human Protein Atlas is 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 [27]. 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 [28].

3. 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 at higher concentration in the diseased state as compared to the healthy reference.

The genes of four proteins that showed elevated concentrations in urine were also differentially expressed on the mRNA level in the transcriptomics dataset. The identified features include the collagen, type XV, alpha 1 (COL15A1), and uromodulin (UMOD), both being upregulated in transcriptomics and proteomics, as well as the prostaglandin D2 synthase 21kDa (PTGDS) and the apolipoprotein A-I (APOA1) that were found at higher concentrations on the protein level in urine although mRNA levels in renal tissue were downregulated (table 2).

Functional overlap

The PANTHER Classification System was used in order to identify enriched biological processes 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 3.

Conjoint pathway analysis

Three pathways could be identified that were 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 4). 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.

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).

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 mRNA molecules of the transcriptomics dataset. 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.

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.

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.

4. 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 corresponding 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 extrinsic 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. Two disease associated features are found in both data sets as upregulated (COL15A1, UMOD), whereas two other jointly found features differ in their regulation. PTGDS is mainly expressed in heart and brain tissue and it's urinary excretion is closely associated with vascular injury and the following damage of renal interstitial regions [29]. Thus, high PTGDS concentrations in urine are 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 [30]. 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.

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 concentrations 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 porteins in urine. For testing this hypothesis we applied 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 [31].

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.

We did further validation of the link of the proteomics data set measured in urine and protein abundance given in kidney compartments. 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, 32].

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).

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

5. REFERENCES

[1] Butte, A., *Nat Rev Drug Discov* 2002, *1*, 951-960.

[2] Hanash, S., *Nature* 2003, *422*, 226-232.

[3] Perco, P., Rapberger, R., Siehs, C., Lukas, A., Oberbauer, R., Mayer, G., Mayer, B., *Electrophoresis* 2006, *27*, 2659-2675.

[4] Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I. F*., et al.*, *Nucleic Acids Res* 2009, *37*, D885-890.

[5] Gygi, S. P., Rochon, Y., Franza, B. R., Aebersold, R., *Mol Cell Biol* 1999, *19*, 1720-1730.

[6] Koji, K., Daisuke, T., Ryoichi, A., Katsutoshi, T., *Genome Letters* 2003, *2*, 139-148.

[7] Lu, P., Vogel, C., Wang, R., Yao, X., Marcotte, E. M., *Nat Biotechnol* 2007, *25*, 117-124.

[8] Shankavaram, U. T., Reinhold, W. C., Nishizuka, S., Major, S., Morita, D., Chary, K. K., Reimers, M. A*., et al.*, *Mol Cancer Ther* 2007, *6*, 820-832.

[9] Rogers, S., Girolami, M., Kolch, W., Waters, K. M., Liu, T., Thrall, B., Wiley, H. S., *Bioinformatics* 2008, *24*, 2894-2900.

[10] Cox, J., Mann, M., *Cell* 2007, *130*, 395-398.

[11] Greenbaum, D., Colangelo, C., Williams, K., Gerstein, M., *Genome Biol* 2003, *4*, 117.

[12] Greenbaum, D., Jansen, R., Gerstein, M., *Bioinformatics* 2002, *18*, 585-596.

[13] Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., Hirakawa, M., *Nucleic Acids Res* 2009.

[14] Brown, K. R., Jurisica, I., *Bioinformatics* 2005, *21*, 2076-2082.

[15] Bernthaler, A., Muhlberger, I., Fechete, R., Perco, P., Lukas, A., Mayer, B., *Mol Biosyst* 2009.

[16] Hallan, S. I., Coresh, J., Astor, B. C., Asberg, A., Powe, N. R., Romundstad, S., Hallan, H. A*., et al.*, *J Am Soc Nephrol* 2006, *17*, 2275-2284.

[17] Rudnicki, M., Eder, S., Perco, P., Enrich, J., Scheiber, K., Koppelstatter, C., Schratzberger, G*., et al.*, *Kidney Int* 2007, *71*, 325-335.

[18] Rudnicki, M., Perco, P., Enrich, J., Eder, S., Heininger, D., Bernthaler, A., Wiesinger, M*., et al.*, *Lab Invest* 2009, *89*, 337-346.

[19] Rossing, K., Mischak, H., Dakna, M., Zurbig, P., Novak, J., Julian, B. A., Good, D. M*., et al.*, *J Am Soc Nephrol* 2008, *19*, 1283-1290.

[20] Schmid, H., Boucherot, A., Yasuda, Y., Henger, A., Brunner, B., Eichinger, F., Nitsche, A.*, et al.*, *Diabetes* 2006, *55*, 2993-3003.

[21] Baelde, H. J., Eikmans, M., Doran, P. P., Lappin, D. W., de Heer, E., Bruijn, J. A., *Am J Kidney Dis* 2004, *43*, 636-650.

[22] Coon, J. J., Zurbig, P., Dakna, M., Dominiczak, A. F., Decramer, S., Fliser, D., Frommberger, M.*, et al.*, *Proteomics Clin Appl* 2008, *2*, 964-973.

[23] Mi, H., Guo, N., Kejariwal, A., Thomas, P. D., *Nucleic Acids Res* 2007, *35*, D247-252.

[24] Dennis, G., Jr., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C., Lempicki, R. A., *Genome Biol* 2003, *4*, P3.

[25] Huang da, W., Sherman, B. T., Lempicki, R. A., *Nat Protoc* 2009, *4*, 44-57.

[26] Marinescu, V. D., Kohane, I. S., Riva, A., *Nucleic Acids Res* 2005, *33*, D91-97.

[27] Berglund, L., Bjorling, E., Oksvold, P., Fagerberg, L., Asplund, A., Szigyarto, C. A., Persson, A.*, et al.*, *Mol Cell Proteomics* 2008, *7*, 2019-2027.

[28] Boguski, M. S., Lowe, T. M., Tolstoshev, C. M., *Nat Genet* 1993, *4*, 332-333.

[29] Yoshikawa, R., Wada, J., Seiki, K., Matsuoka, T., Miyamoto, S., Takahashi, K., Ota, S.*, et al.*, *Diabetes Res Clin Pract* 2007, *76*, 358-367.

[30] Attman, P. O., Knight-Gibson, C., Tavella, M., Samuelsson, O., Alaupovic, P., *Nephrol Dial Transplant* 1998, *13*, 2833-2841.

[31] Perco, P., Pleban, C., Kainz, A., Lukas, A., Mayer, G., Mayer, B., Oberbauer, R., *Eur J Clin Invest* 2006, *36*, 753-763.

[32] Rapberger, R., Perco, P., Sax, C., Pangerl, T., Siehs, C., Pils, D., Bernthaler, A.*, et al.*, *BMC Syst Biol* 2008, *2*, 2.

TABLES

Table 1: Effects influencing omics profile correlation

Intrinsic effects on mRNA and protein abundance	**Extrinsic effects on mRNA and protein abundance**
- transcriptional regulation	- disease mechanisms (e.g. loss of kidney filtration barrier)
- translational regulation	- secretion
- post-translational modifications	- technical bias
- mRNA and protein half-life	

Table 1: Selected intrinsic and extrinsic causes affecting the correlation between mRNA and protein abundance.

Table 2: Direct overlap of differentially regulated omics features

gene symbol	gene name	transcript	protein
COL15A1	collagen, type XV, alpha 1	up	up
UMOD	uromodulin	up	up
PTGDS	prostaglandin D2 synthase 21kDa	down	up
APOA1	apolipoprotein A-I	down	up

Table 2 holds gene symbol and name of features being affected at the transcript or protein level, furthermore providing the direct of regulation.

Table 3: PANTHER biological processes overlap

biological process	**# of genes**	**p-value**	**# of proteins**	**p-value**
Protein metabolism and modification	153	< 0.001	-	-
Blood circulation and gas exchange	-	-	5	< 0.001
Cell structure and motility	78	< 0.001	8	0.0042
Developmental processes	116	< 0.001	-	-
Immunity and defense	80	< 0.001	9	0.0017
Protein modification	70	< 0.001	-	-
Signal transduction	147	< 0.001	-	-
Cell structure	48	< 0.001	8	< 0.001
Cell motility	31	< 0.001	-	-
Intracellular protein traffic	57	< 0.001	-	-
Cell cycle	57	< 0.001	-	-
Cell adhesion	41	< 0.001	5	0.0478
Cell communication	66	< 0.001	-	-
Intracellular signaling cascade	49	< 0.001	-	-
Mesoderm development	36	< 0.001	-	-
Mitosis	28	< 0.001	-	-
Ectoderm development	40	0.0011	-	-
Protein phosphorylation	39	0.0011	-	-
Blood clotting	12	0.0015	-	-
Cell proliferation and differentiation	50	0.0016	-	-
Cell cycle control	28	0.0020	-	-
Neurogenesis	35	0.0028	-	-
Homeostasis	16	0.0034	-	-
Interferon-mediated immunity	9	0.0095	-	-
Angiogenesis	8	0.0255	-	-
Chromosome segregation	12	0.0272	-	-
Apoptosis	27	0.0445	-	-

Table 3 lists biological processes identified as relevant on the basis of given transcriptomics and proteomics data sets. Given is the name of the 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.

Table 4: KEGG pathways overlap

pathway	**# of transcripts**	**p-value**	**# of proteins**	**p-value**
Cell Communication	-	-	6	< 0.001
ECM-receptor interaction	**13**	**< 0.001**	**5**	**< 0.001**
p53 signaling pathway	10	0.01	-	-
Complement and coagulation cascades	**10**	**0.01**	**4**	**< 0.001**
Tight junction	16	0.01	-	-
Regulation of actin cytoskeleton	20	0.02	-	-
Focal adhesion	**18**	**0.05**	**5**	**< 0.001**

Table 4 lists pathway names, 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 feature sets.

FIGURE LEGENDS

Figure 1: KEGG coagulation pathway

Figure 1 displays a schematic representation of the coagulation pathway as provided by the KEGG pathway database. Transcripts are depicted as yellow nodes whereas proteins are given as blue-bordered nodes. Red font indicates upregulated molecules whereas green font indicates downregulated transcripts.

Figure 2: omicsNET dependencies between transcriptomics and proteomics

Figure 2 displays strong dependencies between transcripts and proteins as derived from omicsNET. Grey nodes represent identified proteins while red nodes represent identified transcripts. The two green nodes represent APOA1 and COL15A1 found with differential abundance in both omics profiles.

Figure 3: omicsNET subgraphs of members involved in blood coagulation

Figure 3 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).

Figure 4: omicsNET shortest paths distribution

Figure 4 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.

Figure 5: protein tissue staining

Figure 5 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”.

Figure 1:

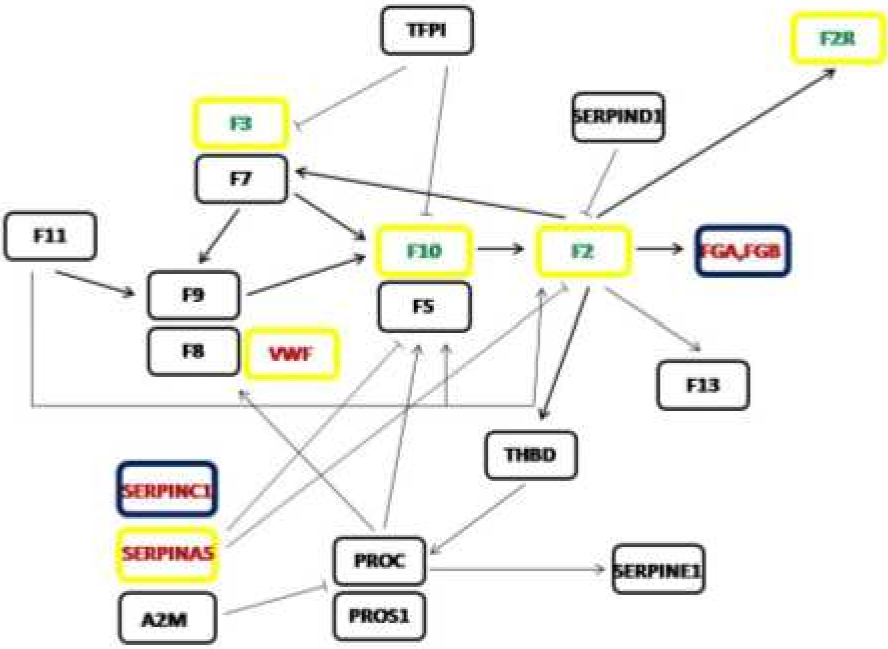

TFPI
F2R
F3
SERPIND1
F7
F11
F10
F2
FGA,FGB
F9
F5
F8
VWF
F13
THBD
SERPINC1
SERPINA5
PROC
SERPINE1
A2M
PROS1

Figure 2:

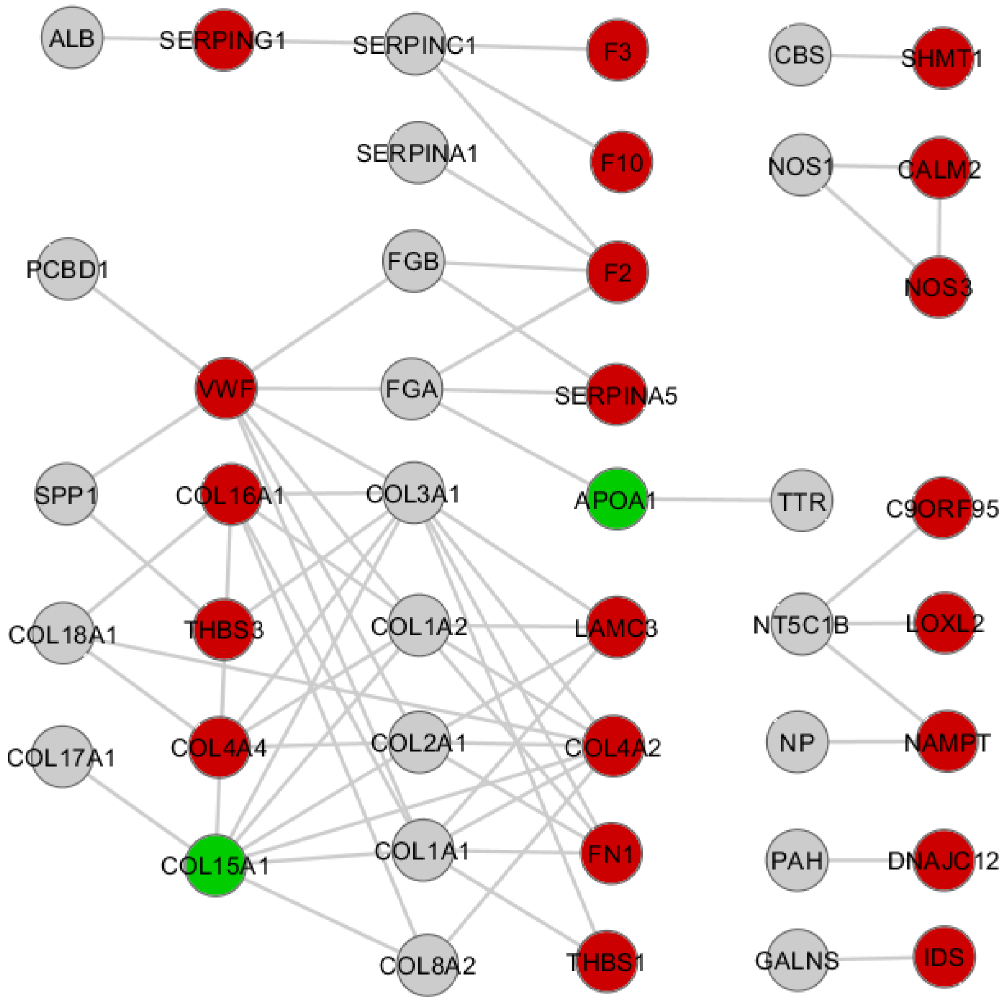
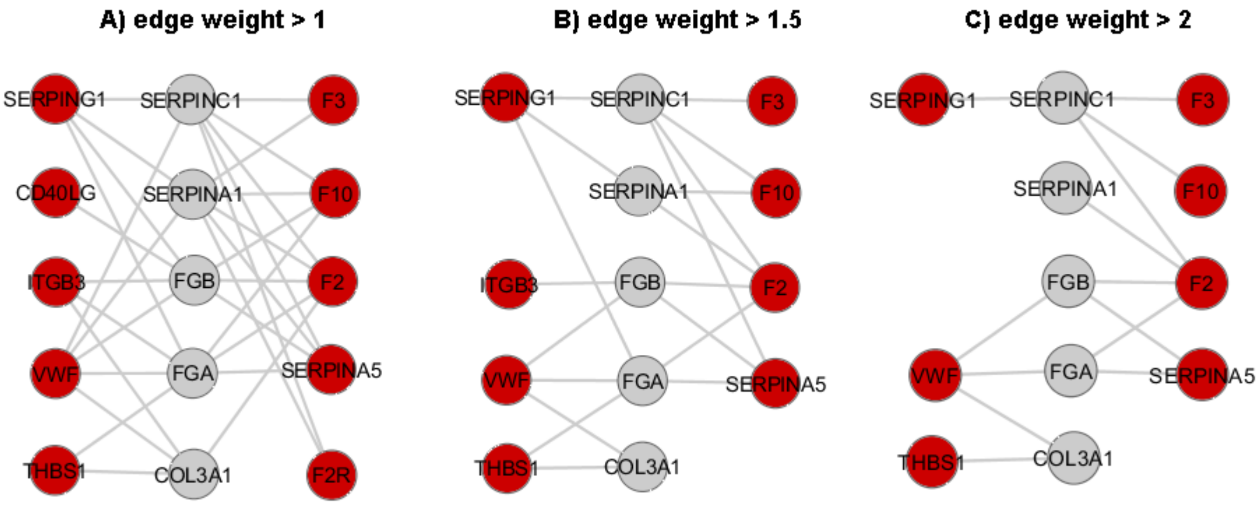

ALB
SERPING1
SERPINC1
F3
CBS
SHMT1
SERPINA1
F10
NOS1
CALM2
PCBD1
FGB
F2
NOS3
VWF
FGA
SERPINA5
SPP1
COL16A1
COL3A1
APOA1
TTR
C9ORF95
COL18A1
THBS3
COL1A2
LAMC3
NT5C1B
LOXL2
COL17A1
COL4A4
COL2A1
COL4A2
NP
NAMPT
COL15A1
COL1A1
FN1
PAH
DNAJC12
COL8A2
THBS1
GALNS
IDS

Figure 3:



A) edge weight > 1
SERPING1
SERPINC1
F3
CD40LG
SERPINA1
F10
ITGB3
FGB
F2
VWF
FGA
SERPINA5
THBS1
COL3A1
F2R
B) edge weight > 1.5
SERPING1
SERPINC1
F3
SERPINA1
F10
ITGB3
FGB
F2
VWF
FGA
SERPINA5
THBS1
COL3A1
C) edge weight > 2
SERPING1
SERPINC1
F3
SERPINA1
F10
FGB
F2
VWF
FGA
SERPINA5
THBS1
COL3A1

Figure 4:

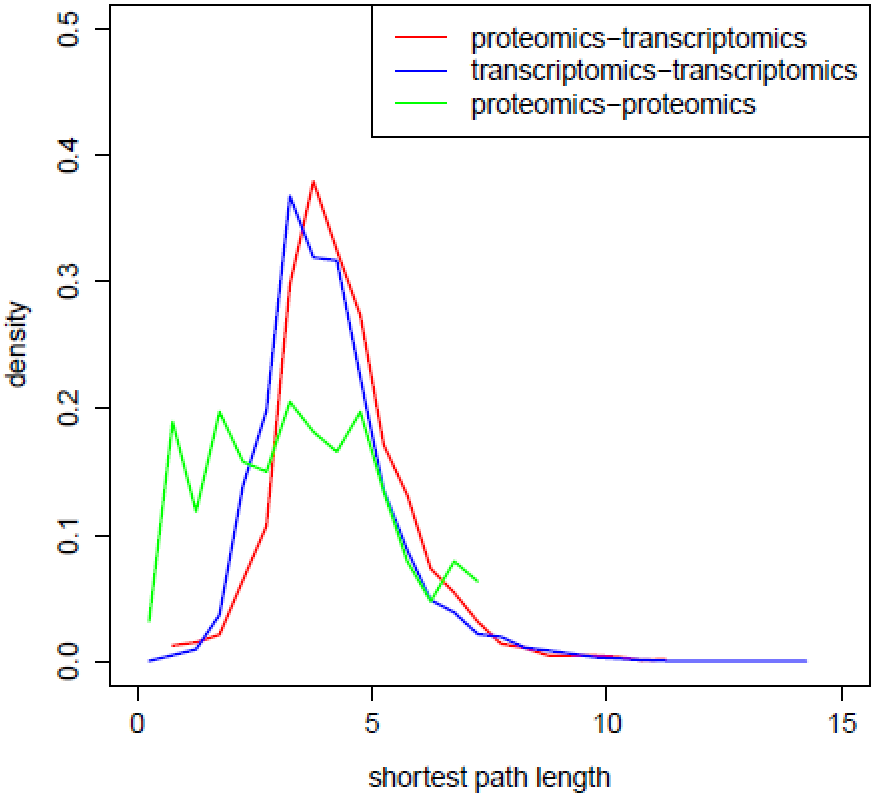

proteomics−transcriptomics
transcriptomics−transcriptomics
proteomics−proteomics
density
0.0
0.1
0.2
0.3
0.4
0.5
0
5
10
15
shortest path length

Figure 5

Gene Symbol	G	T
PGRMC1	░	■
SPP1	░	■
PTGDS	□	■
TTR	□	■
UMOD	□	■
ALB	▩	▩
B2M	▩	▩
FGA	▩	▩
ORM1	░	▩
AHSG	□	▩
FXYD2	□	▩
SERPINC1	□	▩
APOA1	▩	░
COL15A1	░	░
COL3A1	□	░
CSTB	□	░
FGB	□	░
PIGR	□	░
SERPINA1	□	░
COL1A1	■	□
CD99	▩	□
A1BG	□	□
COL18A1	□	□
COL2A1	□	□
PCSK1N	□	□

staining intensity	
■	strong
▩	moderate
░	weak
□	negative
☒	not available

Gene Symbol	G	T
COL1A2	☒	☒
COL8A2	☒	☒
COL17A1	☒	☒
HBA1	☒	☒
HBA2	☒	☒
HBB	☒	☒
IGL@	☒	☒
IGLC2	☒	☒
IGLV2-14	☒	☒
PSORS1C2	☒	☒
VGF	☒	☒
ZNF653	☒	☒