Aus der Klinik für Nephrologie der Medizinischen Fakultät Charité –

Universitätsmedizin Berlin

DISSERTATION

Bioinformatics Modeling of Proteomics changes in Muscle Invasive Bladder Cancer

zur Erlangung des akademischen Grades

Doctor rerum medicinalium (Dr. rer. medic.)

vorgelegt der Medizinischen Fakultät

Charité – Universitätsmedizin Berlin

von

Akshay Bhat aus Mumbai / Indien

Datum der Promotion: 09/09/2016

Table of Content

LIST OF ABBREVIATIONS

Abbreviation	Meaning
BC	Bladder Cancer
BCG	Bacillus Calmette-Guerin
CIS	Carcinoma in-situ
GEO	Gene Expression Omnibus
GO	Gene Ontology
KEGG	Kyoto Encyclopedia of Genes and Genomes
MIBC	Muscle Invasive Bladder Cancer
OMIM	Online Mendelian Inheritance in Man
PE	Pathway Enrichment
PPI	Protein-Protein Interaction
RDBMS	Relational Database Management System
ROR	Ruby on Rails
TCGA	The Cancer Genome Atlas
TNM	Tumor-Node-Metastasis

1. Abstract (German):

Blasentumore können in zwei Phänotypen eingeteilt werden: dem muskelinvasiven beziehungsweise dem nicht-*muskelinvasiven* Blasentumor. Dernicht-*muskelinvasive* Blasentumor hat durch eine Tumorsektion und einer intravesikale Therapie gute Prognosen. Der *muskelinvasive* Blasentumor hat jedoch trotz Zystektomie und Cisplatin-basierter Chemotherapie weiterhin eine schlechte Therapieprognose. Für eine erfolgreiche Therapie ist die Identifizierung der zugrundeliegenden Stoffwechselwege des jeweiligen Phänotyps umso wichtiger. Um hierzu einen Beitrag zu leisten, wurden im Rahmen der vorliegenden Dissertation drei Studien durchgeführt. Im ersten Teil der Dissertation wurden Blasentumore aufgrund von verfügbaren Omics-Profilen und aufgrund von Literaturdaten charakterisiert. Im zweiten Teil der Arbeit wurden zur Blasentumor-Charakterisierung bioinformatische Datenbanken analysiert und im dritten Teil der Dissertation wurden Signalwege der Blasentumorgenese und –Progression anhand eines bioinformatischen Ansatzes analysiert. Die öffentlich zugänglichen Omics-Datenbanken und Datenbanken molekularer Merkmale des *muskelinvasiven* Blasentumors wurde mit dem Ziel analysiert, neue Zielgene der Erkrankung zu identifizieren. Die betreffenden Merkmale wurden in ein Protein-Protein-Interaktionsnetz integriert, um hierdurch relevant Signalwege ausfindig zu machen. Hierzu wurde die Software-Werkzeuge ClueGo und ImPAla verwendet und gegenzeitig verglichen. Die resultierenden Signalwege wurden statistisch bewertet. Mit Hilfe dieses Ansatzes konnten insgesamt 14 Signalwege des Blasentumors identifiziert werden. Drei dieser Stoffwechselwege waren bisher noch nicht im Kontext des Blasentumors beschrieben worden; hierbei handelte es sich um Aktin- Zytoskeletton, den Neurotrophin-Signalweg sowie der Endozytose.

Im zweiten Teil der Dissertation wurden durch die Verwendung von ClueGo 292 Signalwege und 471 Signalwege durch den Einsatz des ImPAla- Softwaretools extrahiert. Es zeigt sich

eine Übereinstimmung von 152 Signalwegen bei diesen Ansätzen. 137 der mittels ClueGo extrahierten Signalwege waren mit den 251 ImPAla Signalwege vergleichbar. Im dritten Teil der Dissertation führte die bioinformatische Analyse von proteomischen Daten des Urins zur Identifizierung eines Signalweges, der für zugrunde liegende Alterungsprozesse relevant zu sein scheint. Insgesamt konnte im Rahmen der Dissertation gezeigt werden, dass ein Vielzahl sehr unterschiedlicher Signalwege in der Entwicklung des *muskelinvasiven* Blasentumor involviert sind. Die im Rahmen der Dissertation neu-identifizierten Signale sind im Rahmen von Folgeuntersuchungen zu validieren, um deren Bedeutung für den *muskelinvasiven* Blasentumor zu bestimmen. Das Software-Werkzeug ClueGo ist –zumindest im Kontext des *muskelinvasiven* Blasentumors- dem ImPAla-Ansatz aufgrund der geringeren Redundanz und des biologischen Informationsgehalts vorzuziehen. Darüber hinaus scheint die Analyse molekularer Signalwege anhand von peptidomischen Daten und Techniken geeignet zu sein, zusätzliche Informationen über Alterungsprozesse zu bieten.

1. Abstract (English)

BACKGROUND: Bladder Cancer (BC) has two distinct phenotypes. Non-muscle invasive BC has good prognosis and is treated by tumor resection and intravesical therapy, whereas muscle-invasive BC has poor prognosis and requires radical cystectomy combined with cisplatin-based chemotherapy. High-throughput sequencing technologies allow identification of individual molecular signatures to characterize the invasive phenotype.

OBJECTIVE: Based on this background the objective of this thesis comprises of three parts. The first aim of the thesis was to characterize muscle-invasive BC on a molecular level by incorporating signatures from literature and omics profiles. The second aim was to evaluate the performance of pathway-enrichment obtained from two bioinformatics tools ImPAla and ClueGO. The third aim shows the use of bioinformatics in order to identify altered pathways relevant to ageing.

MATERIALS: Public domain -omics signatures and molecular features associated to muscle-invasive BC were derived from literature mining to provide protein-coding genes. These features were integrated in a protein-interaction network to obtain functional pathways relevant to the phenotype. Pathway-enrichment was performed using ClueGO and ImPAla tools. The resulting pathway terms were filtered according to criterion: multiple comparison corrected p-value <0.05.

RESULTS: In the first part of the thesis, the protein-interactions and pathway-enrichment yielded 14 significant pathway terms. Three pathway terms were not previously reported in muscle-invasive BC. The novel disease-associated pathways were regulation of actin-cytoskeleton, neurotrophin-signalling pathway and endocytosis. In the second part, 292 pathways were obtained from ClueGO and 471 pathways from ImPAla software. Comparison of the results obtained by the two applications yielded 152 pathway-terms with the same pathway name. 137 ClueGO pathway-terms were similar to 251 ImPAla pathways. In the last part, the results from a bioinformatics analysis of urinary-peptidomics data discovered a pathway-term “degradation of insulin-like growth factor-binding proteins” that was unique in the context of pathological ageing.

CONCLUSIONS: The results of this thesis suggest that there is a complex interplay between pathways characterizing the muscle-invasive phenotype of BC. Further experimental validation of the three novel pathways with respect to progression and treatment response is indicated. In addition, the comparison of two prominent pathway enrichment tools ClueGO and ImPAla showed that ClueGO has better performance than ImPAla in pathway-enrichment analysis since the output is less redundant and contains all the biologically significant information. Lastly, molecular pathways enriched in normal and pathological ageing demonstrate that with the help of appropriate peptidomics technologies, urine could be used as a useful source of information in ageing research.

2. Introduction

Urothelial bladder carcinoma is a common malignancy of the urinary tract system and comprises of two distinct clinical phenotypes, non muscle-invasive bladder cancer and muscle-invasive tumors. Muscle-invasive bladder tumors are treated with pre-operative (neoadjuvant) cis-platin-based chemotherapy followed by removal of the bladder, named as "radical cystectomy" [1]. However, since a significant number of patients do not respond to chemotherapy treatment, a detailed investigation of the tumor molecular signature is required to select responsive patients for bladder cancer treatment [2]. A better understanding of muscle invasive bladder cancer might be achieved by combining information obtained from individual biomarkers measured at the DNA, RNA and/or protein levels[3]. Along this background, the comprehensive characterization of altered molecular pathways provides significant clinical relevance in order to choose optimum diagnosis and treatment regimens for bladder cancer patients.

Variable for individual bladder cancer patients, initial symptoms include hematuria and flank pain [4,5]. Cystoscopy is the gold standard diagnostic procedure with a reported sensitivity of 62–84% and specificity of 43–98%. This wide variability in sensitivity and specificity indicates a significant inter-operator variability[6]. Non muscle-invasive bladder cancer comprises of distinct forms [7–9]:

a. Ta stage - the cancer is just in the innermost layer of the bladder lining.
b. T1 stage - the cancer has started to grow into the connective tissue beneath the bladder lining
c. carcinoma in-situ (CIS) - very early, high grade, cancer cells are only in the innermost layer of the bladder lining [9].

Papillary tumors that are confined to the mucosa and that invade the lamina propria of the bladder are classified as stages Ta and T1 according to the Tumor-Node-Metastasis (TNM) classification system [10]. The papillary carcinoma (Ta and T1) phenotype has a tendency to recur locally and it rarely invades the bladder basement membrane or metastasizes to adjacent organs. However, the flat carcinoma in situ (CIS) is often multifocal and is a dangerous lesion with a high tendency for bladder muscle invasion and metastasis[11]. Treatment of non-muscle invasive bladder carcinoma (Ta, T1 and CIS) involves endoscopic transurethral resection of visible tumors followed by adjuvant treatment with intravesical instillation therapy (Mitomycin/Epirubicin or Bacillus Calmette-Guerin (BCG)) depending on the estimated risk for progression. Irrespective of aggressive treatment and vigorous follow-up, 70% of these tumors recur, and 25% of high-grade non-muscle invasive cancers (CIS)

progress into invasive phenotypes [12,13].

Muscle invasive bladder cancers are advanced cancer stages and are classified as “T2 – T4” according to the Tumor-Node-Metastasis classification system.

a) T2 stage – is when the cancer has spread into the muscle layer of the bladder
b) T3 stage – is when the tumor has grown through the muscle layer
c) T4 stage – is when the cancer has spread into the prostrate, uterus or vagina, or into the wall of the pelvis or abdomen

Furthermore, muscle invasive bladder tumors are also distinguished into three distinct molecular subtypes that have widely variable clinical outcomes and responses to conventional chemotherapy treatments:

a. Basal subtypes
b. Luminal subtypes
c. “p53-like" tumors [14]

The basal muscle invasive bladder cancer subtypes are susceptible to have more invasive and metastatic disease at initial diagnosis and are associated with shorter disease-specific and overall survival. The biomarkers for the basal muscle invasive bladder cancer subtype include CD44 antigen (CD44), Keratin, type II cytoskeletal 5 (KRT5), Keratin, type II cytoskeletal 72 (KRT72), Keratin, type I cytoskeletal 14 (KRT14) and Cadherin-3 (CDH3). The luminal muscle invasive bladder cancer subtypes are enriched with activating fibroblast growth factor receptor 3 (FGFR3) and human epidermal growth factor receptor 3 (ERBB3) mutations and Receptor tyrosine-protein kinase erbB-2 (ERBB2) amplifications, and the gene expression profiles are controlled by peroxisome proliferator activator receptor γ (PPARγ) and oestrogen receptor activation. The wild-type p53 is required for DNA damage induced apoptosis and is a central tenet in cancer biology [15]. Therefore, it is interesting that the de-novo and induced chemoresistance in muscle invasive bladder cancers was associated with wild-type p53 gene expression signatures. Nevertheless, TP53 mutation frequencies are similar in all three subtypes of muscle invasive bladder cancers (basal, luminal and p53-like), indicating that wild-type p53 was not responsible for the baseline and chemotherapy induced p53-like gene expression. Hence, it is proposed that ‘‘p53-like tumors’’ as measured by mRNA expression would be a more accurate predictor of de-novo and induced muscle invasive bladder cancers chemoresistance than would analyses of TP53 mutational status[16]. The determination of the molecular basis of these p53-like signatures is not yet defined and that could overcome de novo and/ or prevent acquired chemoresistance [14].

The most important point for the planning of radical cystectomy in bladder cancer

tumors is the depth of invasion or muscular involvement (T category, P stage)[17]. Various treatments have proved useful for disease control in some patients with regional bladder cancer but the most common procedure for this situation is radical cystectomy[18]. A recent report highlights significant errors in clinical staging of patients with bladder cancer that underwent radical cystectomy [19]. This staging error of cystectomy in the bladder tumors may cause very important mistakes on the decision for radical surgery treatment. Hence, obtaining an accurate staging diagnosis is particularly crucial in patient selection for surgical treatments (i.e. cystectomy) and the choice of chemotherapy. Due to the invasive procedure of cystoscopy and in order to improve accuracy in the phenotype detection, blood or urine biomarkers could support clinical assessment [20].

High-throughput experimental platform technologies range from genomic sequencing to epigenomic, transcriptomic, proteomic and metabolomic profiling in order to characterize the molecular aspects of individual clinical phenotypes [21–28]. Genomic sequencing comprises of applying recombinant DNA, DNA sequencing methods, and bioinformatics to sequence, assemble, and analyze the function and structure of "genomes" (i.e. the complete set of DNA within a single cell of an organism). Epigenomic sequencing is the determination of key functional elements that regulate gene expression in a cell – Epigenomes provide information about the patterns in which structures such as methyl groups tag DNA and histones (the proteins around which DNA is packaged to form the chromatin), and about interactions between distant sections of chromatin). Transcriptomics is the sequencing and quantification of transcripts – mRNA and microRNA or miRNA. Proteomics is the sequencing and quantification of the proteome and peptidome, and metabolomics is the technique of identifying and quantifying of metabolites. These techniques provide datasets that comprise of DNA-mutations, DNA-methylations, mRNAs, miRNA, proteins, peptides and metabolites. The advent of these approaches that generate a comprehensive view of the molecular landscape for a biological sample has introduced a paradigm shift in the way diseases are perceived [21,22,29].

A variety of datasets for such molecular characterizations have become available that are stored in public databases, for e.g. in Array Express[30] or Gene Expression Omnibus (GEO)[31], which is a database that stores mRNA and miRNA datasets from transcriptomics experiments, Human Proteinpedia is a public repository that provides information on proteomics datasets[32,33], Human Protein Atlas (http://www.proteinatlas.org/), which is an online portal that contains information on immunohistochemically validated proteins, or large data consolidation resources such as GeneCards[34] that provides information for genomic,

proteomic, transcriptomic, genetic and functional information on all known and predicted human genes. This database aims to provide a quick overview of the current available biomedical information about the searched gene, including the human genes, the encoded proteins, and the relevant diseases.

In regard to disease specific omics data, valuable general sources in oncology include The Cancer Genome Atlas (TCGA) (http://cancergenome.nih.gov/), Oncomine [35], and Online Mendelian Inheritance in Man (OMIM) [36]. The Cancer Genome Atlas oncology portal currently lists single nucleotide polymorphism, methylation data, mutations, mRNAs, miRNAs and proteins relevant to bladder cancer. A recent report presents a systems biology approach for the analysis of the muscle invasive bladder cancer dataset contained in The Cancer Genome Atlas [37]. Another database for bladder cancer that provides molecular features in regard to miRNAs identified in literature is also available [38,39]. In addition, a user-friendly analysis tool is also available and allows the evaluation of gene expression profiles determined by microarray studies across bladder cancer patients [40].

Though omics profiling has provided an abundance of data, technical boundaries involving incompleteness of the individual molecular datasets together with the static representation of cellular activity limit the insights on molecular processes and their interaction dynamics [41–43].

A large number of biological pathway analysis tools are available, including KEGG [44], PANTHER [45], REACTOME [46] and AmiGO [47] described in PathGuide (http://www.pathguide.org/), and allow detection of significant metabolic and signaling pathways. Albeit there are several well curated and reliable pathway database resources [48], significant efforts have been taken to expand biological pathway coverage beyond any single pathway data source. This is frequently carried out by integrating different pathway sources to build high quality integrative pathway models. However, biological data integration from heterogeneous sources has been challenging due to variability at the syntactic and semantic level. Syntactic variability is due to heterogeneity of molecular feature and pathway data formats, representation schemas and retrieval methods. Semantic variability is due to incompatible pathway names, signaling event representations and molecular identifiers. For example, different pathway databases may choose to provide information on post-translation modifications, interacting proteins within a complex, or cellular location. Hence all these limitations have inhibited the growth of high quality integrative pathway models [49–51].

Previous omics studies report biomarkers associated with bladder cancer, and therapeutic targets that could allow development of personalized therapies [52–56]. However, the information gathered from these large number of omics experiments is not fully exploited, as

the datasets generated are either scattered in many publications and databases or held in supplementary data files.

Therefore, the aim of this thesis was primarily to characterize muscle invasive bladder carcinoma on a molecular level by incorporating scientific literature and omics data. In addition, the objective was to evaluate the performance of pathway enrichment analysis obtained from two bioinformatics tools ImPAla and ClueGO. Thirdly, the goal was to use bioinformatics and systems biology approaches in order to identify significant molecular pathways in age-associated diseases.

3. Methods

3.1 Data sources for characterizing bladder cancer pathophysiology

In order to retrieve molecular features associated with muscle invasive bladder cancer, "National Center for Biotechnology Information" (NCBI) PubMed, Web of Science, Google Scholar and the omics repositories Gene Expression Omnibus (GEO)[31] and ArrayExpress[30] were manually queried. Since the study involved the molecular characterization of muscle invasive bladder cancer, the criteria for selecting keywords depended specifically in regard to the muscle invasive phenotype. The keywords for the literature search included "bladder OR urothelial OR transitional cell" AND "neoplasm OR tumor OR carcinoma" AND "muscle" AND "invas* OR aggress* OR progress* OR inflammation" (Database version of June, 2015). The list of publications relevant to muscle invasion in bladder cancer was isolated from the complete list of retrieved papers. Publications were further screened for adequacy in sample size (at least 50 samples included in study design), magnitude of differential abundance (>2-fold change for proteomics, transcriptomics, metabolomics and miRNAs), False Discovery Rate <0.1 for mutations, p-values<0.05 for methylation and –omics studies, in addition to the specific phenotypic conditions; T2a/b, T3a/b, T4a/b. The muscle invasive bladder cancer specific molecular features retrieved from the publications comprised of various sources such as DNA-mutations, DNA-methylation, mRNAs, miRNAs, proteins (immunohistochemistry validations and proteomics) and metabolites. The features were then combined for further systems biology analysis.

3.2 Protein-protein interactions

In order to retrieve protein-protein interaction information for the muscle invasive bladder cancer associated proteins, protein-protein interaction databases including IntAct[57], BioGRID[58], STRING[59] and Reactome[46] were queried. By downloading the protein interaction information contained in each database, an integrated database was developed in order to contain all available non-redundant human protein-protein interaction information. This unique human protein list along with the protein-protein interactions information were then downloaded into the Cytoscape[60,61] software to yield the human interactome based on experimental evidence. The proteins relative to muscle invasive bladder cancer were then selected from this human interactome and were put on a separate list. Muscle invasive bladder cancer proteins that had at least one binding partner in the list of muscle invasive bladder cancer specific proteins were retained in order to generate the muscle invasive bladder cancer specific interactome.

3.3 Pathway enrichment

To retrieve molecular pathway information for muscle invasive bladder cancer, proteins from the muscle invasive bladder cancer interactome were subjected to pathway enrichment analysis. This analysis used two additional sub-applications from Cytoscape; ClueGO[62] and CluePedia[63]. The statistical criterion used in generating molecular pathways included a two-sided hypergeometry test. Information from pathway databases such as Kyoto Encyclopedia of Genes and Genomes [44] and Reactome[46] databases was used in retrieving significant pathways associated to muscle invasive bladder cancer with a Bonferroni corrected p-value<0.05. In addition, the list of pathways was inspected manually and redundant pathway-terms were combined hereby. The filtered list of pathway-terms was then divided into previously known pathways and novel findings in the context of muscle invasive bladder carcinoma.

3.4 Comparison of pathway enrichment tools ClueGO and ImPAla

To evaluate the performance of pathway enrichment, ClueGO and ImPAla enrichment tools were compared. ClueGO provides an advantage to perform cluster comparisons for pathway enrichment and allows the option to separately input up and down regulated molecules in the software. In addition, ClueGO provides an optional redundancy reduction feature ("Fusion") to assess Gene Ontology (GO) terms that share similar associated features in a parent–child relation. This option was selected in the ClueGO pathway enrichment analysis to eliminate the redundant pathway terms. In contrast, ImPAla does not provide an option of redundancy reduction for pathway terms. The pathway databases selected for enrichment were KEGG. The statistical selection criterion taken into account for the enrichment analysis was the corrected for multiple comparisons p-value<0.05. The overlap assessment between the pathway outputs was performed manually.

4. Results

4.1 Publication 1: Protein Interactome for Muscle Invasive Bladder Cancer.

Akshay Bhat, Andreas Heinzel, Bernd Mayer, Paul Perco, Irmgard Mühlberger, Holger Husi, Axel S. Merseburger, Joost P Schanstra, Jerome Zoidakis, Antonia Vlahou, Harald Mischak, *Vera Jankowski*. PLoS One. 2015.

In the present study the bioinformatics model of proteomic changes in bladder cancer involved integrating available public domain data sets from PubMed, Google scholar and Web of science in the context of bladder muscle-invasive carcinoma (**Figure 1**).

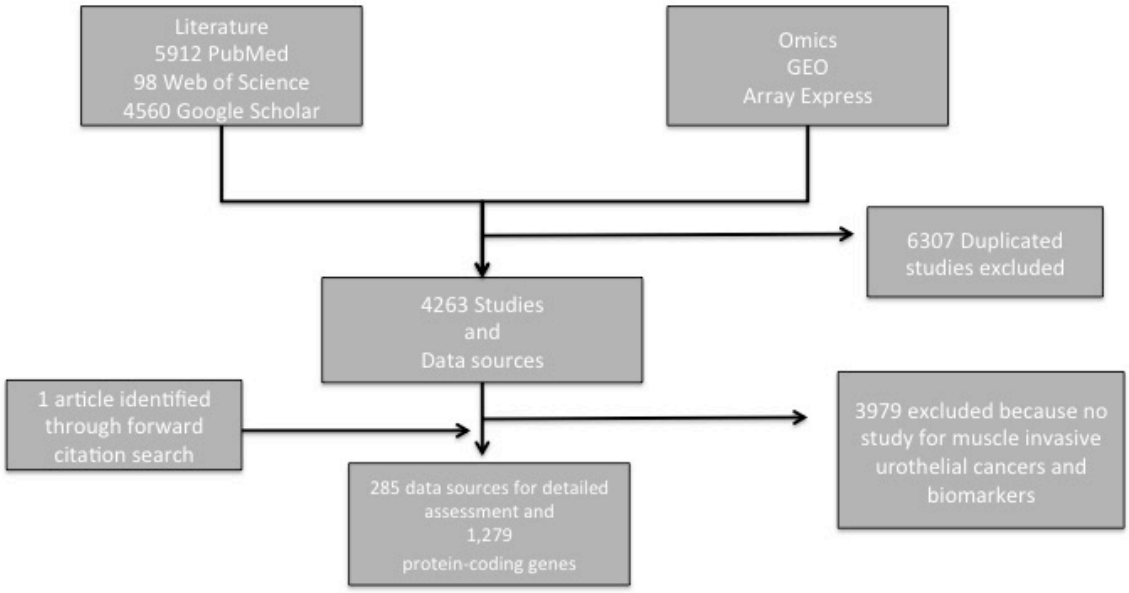

Figure 1. Data assembly workflow. PubMed, Google Scholar and Web of Science literature analysis and Omics data source screening for the systems based analysis in muscle invasive bladder cancer.

In the first step of this analysis, the data collected was filtered using statistical measurements to include fold-change values, p-values and sample size for the specific phenotype of muscle invasive bladder cancer. The molecular features were then incorporated into systems biology tools to model protein-protein interaction networks, and further mapping them to biological molecular pathways.

The results revealed fifteen pathways as being affected in the progressive disease. Eleven from these pathways were reported previously and four pathways were novel findings in the context of muscle invasive bladder cancer (**Figure 2**). The fact that the majority of pathways identified by our analysis are involved in muscle invasive bladder cancer supports the validity of our approach. Moreover, the four novel pathways revealed by our analysis could be validated experimentally and offer new targets for biomarker discovery or therapy of muscle invasive bladder cancer.

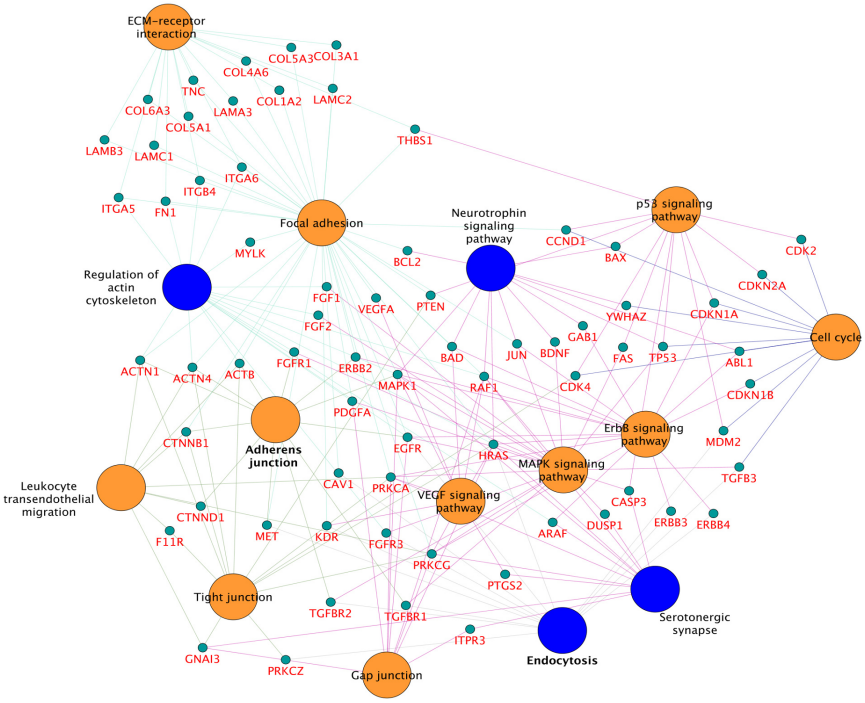

Figure 2. Muscle Invasive Bladder carcinoma interactome.
Nodes (circles) in orange denote pathways identified as relevant in both literature and enrichment analysis, nodes in blue depicts pathways of relevance according to enrichment analysis.

Table 1. Abbreviations for protein coding genes described in Figure 2.

LAMB3 – Laminin subunit beta-3	LAMA3 - Laminin subunit alpha-3
COL6A3 – Collagen alpha-3(VI) chain	COL1A2 - Collagen alpha-2(I) chain
TNC - Tenascin	LAMC2 - Laminin subunit gamma-2
COL4A6 - Collagen alpha-6(IV) chain	ITGA5 - Integrin alpha-5
COL5A3 - Collagen alpha-3(V) chain	FN1 - Fibronectin
COL3A1 - Collagen alpha-1(III) chain	ITGB4 - Integrin beta-4
LAMC1 - Laminin subunit gamma-1	ITGA6 - Integrin alpha-6
COL5A1 - Collagen alpha-1(V) chain	MYLK - Myosin light chain kinase, smooth muscle
THBS1 - Thrombospondin-1	ACTN1 - Alpha-actinin-1
ACTN4 - Alpha-actinin-4	ACTB - Actin, cytoplasmic 1
FGFR1 - Fibroblast growth factor receptor 1	FGF2 - Fibroblast growth factor 2

FGF1 - Fibroblast growth factor 1

BCL2 - Apoptosis regulator Bcl-2

CCND1 - G1/S-specific cyclin-D1

CTNNB1 - Catenin beta-1

PDGFA - Platelet-derived growth factor subunit A

BAD - Bcl2-associated agonist of cell death

F11R - Junctional adhesion molecule A

KDR - Vascular endothelial growth factor receptor 2

EGFR - Epidermal growth factor receptor

JUN - Transcription factor AP-1

GAB1 - GRB2-associated-binding protein 1

CDKN1A - Cyclin-dependent kinase inhibitor 1

CDK2 - Cyclin-dependent kinase 2

TGFBR2 - TGF-beta receptor type-2

PRKCA - Protein kinase C alpha type

CDK4 - Cyclin-dependent kinase 4

TP53 - Cellular tumor antigen p53

CDKN1B - Cyclin-dependent kinase inhibitor 1B

TGFB3 - Transforming growth factor beta-3

ERBB3 - Receptor tyrosine-protein kinase erbB-3

DUSP1 - Dual specificity protein phosphatase 1

PTGS2 - Prostaglandin G/H synthase 2

TGFBR1 - TGF-beta receptor type-1

PRKCZ - Protein kinase C zeta type

PRKCG - Protein kinase C gamma type

VEGFA - Vascular endothelial growth factor

PTEN - Phosphatidylinositol 3,4,5-trisphosphate 3-phosphatase and dual-specificity protein phosphatase PTEN

BAX - Apoptosis regulator BAX

ERBB2 - Receptor tyrosine-protein kinase erbB-2

MAPK1 - Mitogen-activated protein kinase 1

CTNND1 - Catenin delta-1

MET - Hepatocyte growth factor receptor

CAV1 - Caveolin-1

RAF1 - RAF proto-oncogene serine/threonine-protein kinase

BDNF - Brain-derived neurotrophic factor

YWHAZ - 14-3-3 protein zeta/delta

CDKN2A - Cyclin-dependent kinase inhibitor 2A

GNAI3 - Guanine nucleotide-binding protein G(k) subunit alpha

FGFR3 - Fibroblast growth factor receptor 3

HRAS - GTPase Hras

FAS - Tumor necrosis factor receptor superfamily member 6

ABL1 - Tyrosine-protein kinase ABL1

MDM2 - E3 ubiquitin-protein ligase Mdm2

ERBB4 - Receptor tyrosine-protein kinase erbB-4

CASP3 - Caspase-3

ARAF - Serine/threonine-protein kinase A-Raf

PRKCG - Protein kinase C gamma type

ITPR3 - Inositol 1,4,5-trisphosphate receptor type 3

TGFBR2 - TGF-beta receptor type-2

4.2 Publication 2: Comparison of ClueGO and ImPAla for integrated pathway enrichment analysis.

Akshay Bhat, *Vera Jankowski*, Antonia Vlahou, Harald Mischak, Jerome Zoidakis. Accepted in Jacobs Journal of Bioinformatics and Proteomics –[Epub ahead of print]

In this study, the total number of Kyoto Encyclopaedia of Genes and Genomes (KEGG) and Reactome pathway terms obtained from ClueGO was 292. ImPAla produced 471 pathways (**Table 1**).

Table 2. General information for the results obtained from the pathway enrichment analysis.

Software	Availability	User input	p-value correction method	Total pathway output
ClueGO	Cytoscape plugin	435 entries	Bonferroni	292
ImPAla	Web-based	435 entries	Benjamini Hochberg	471

By comparing the pathway results, 152 pathway terms exactly overlapped in ClueGO and ImPAla. 137 pathway terms from ClueGO were highly similar to 251 ImPAla pathway terms. Therefore, the total calculated overlap of pathways between the two tools equalled to 289 ClueGO pathways that correspond to 403 ImPAla enriched pathways. In addition, the software also produced unique pathway terms. There were 3 unique pathways from the total 292 ClueGO pathway terms whereas 68 pathways were unique from ImPAla. Both the enrichment tools yielded redundancy in the output results, however results from ImPAla were characterized by higher redundancies in pathway terms (for e.g. the pathway terms "DNA replication", "synthesis of DNA"). Moreover, from the unique set of 68 ImPAla pathway terms, 12 pathway terms were not related in the context of bladder cancer. Some of these pathways include alcoholism, amphetamine addiction, inflammatory bowel disease (IBD), malaria, viral myocarditis and prion diseases. On the contrary, the 3 unique pathways obtained by ClueGO were relevant to bladder cancer. It was also noted that the overlapping pathway terms from ImPAla and ClueGO contained pathway names that are not relevant in the context of bladder cancer. These common terms totalled to 34 ImPAla and 30 ClueGO pathway terms. The common pathway terms included oocyte meiosis, tuberculosis, type II diabetes mellitus, circadian clock and shigellosis. The comparison of significant overlapping pathways obtained from ClueGO and ImPAla is represented as a Venn diagram in (**Figure 3**).

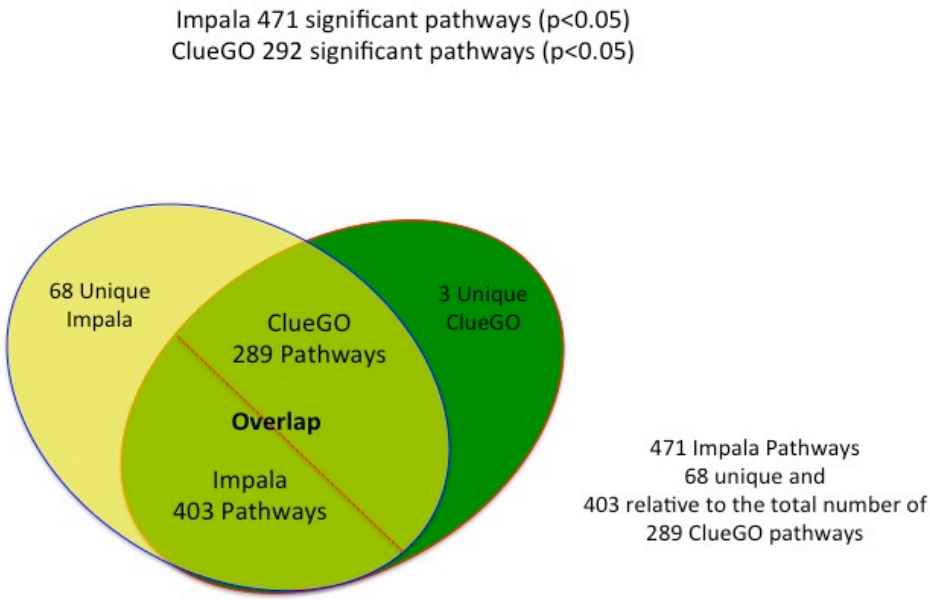

Figure 3. Venn Diagram represents the overlap of pathway terms between ClueGO and ImPAla software. All pathways enriched are selected based on p-value < 0.05

4.3 Publication 3: Identification of urinary age-specific peptides in a healthy population.

Esther Nkuipou-Kenfack, **Akshay Bhat**, *Vera Jankowski*, Julie Klein, William Mullen, Antonia Vlahou, Thomas Koeck, Mohammed Dakna, Joost P. Schanstra, Petra Zürbig, Andreas Pich, Harald Mischak. DOI: 10.18632/oncotarget.5896

In the last section of the thesis, the bioinformatics approach used in the molecular characterization of bladder cancer muscle invasion was applied to a peptidomics dataset relevant to ageing associated disorders.

Ageing is a complex systemic process and "omics" approaches aiming at the study of multiple features simultaneously have been applied to unravel novel underlying molecular processes[64]. Proteomics studies confirmed that oxidative stress occurs ubiquitously during ageing[65]. However, a shortcoming in most of these studies was the use of animal models. The scarcity of human subjects can be largely attributed to the inability in obtaining appropriate tissue samples. Thus, a way forward in ageing research could be the investigation of readily available body fluids.

In this study, a small-scale urinary peptidome of 324 healthy individuals was investigated. The patients aged between 2 to 73 years and showed the feasibility to obtain high-resolution molecular information readily available from body fluids such as urine[66]. Subsequently, the urinary peptidome profiles of 11,560 individuals were investigated in an attempt to identify specific ageing-associated alterations and to elucidate pathological derailment in normal ageing. The results obtained showed perturbations mainly in collagen homeostasis, trafficking of toll-like receptors and endosomal pathways that were significantly associated to the healthy ageing group. Moreover, degradation of insulin-like growth factor-binding proteins was a unique identification deregulated in pathological ageing cohorts (**Figure 4a and 4b**).

Legends for figure 4a and 4b

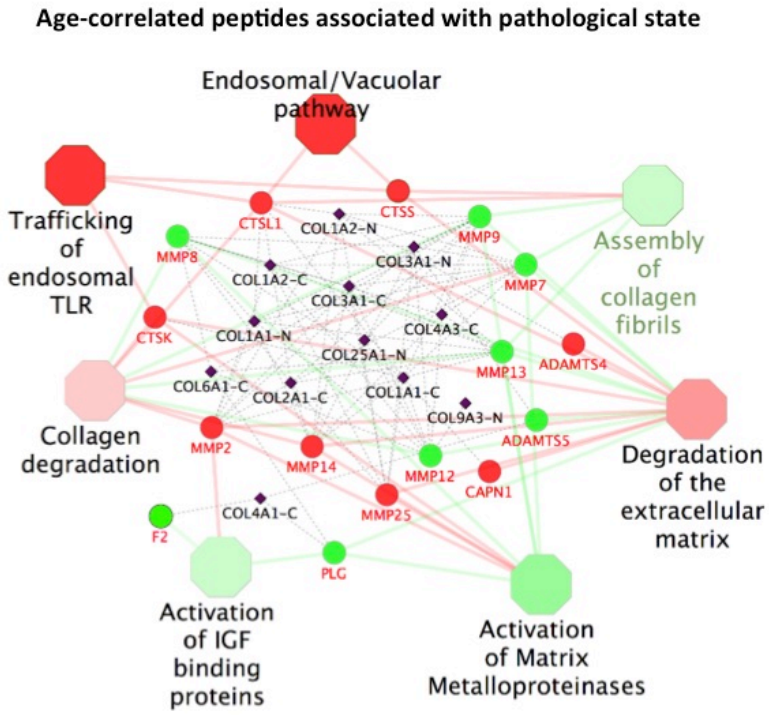

Figure 4a: Molecular pathways associated with pathological ageing. The network represents each pathway as individual octagonal node, while the circled nodes represent the predicted proteases that were targeted from the identified urinary peptides denoted in purple diamond nodes. The edges (links) between pathways denote an approximation of biological interaction between the pathways based on the cross-pathway feature overlap.

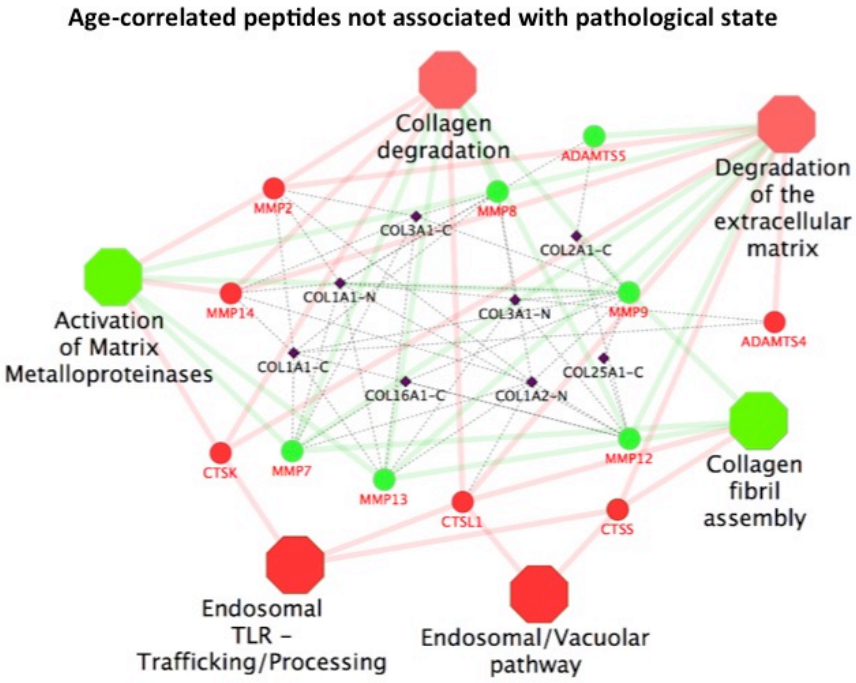

Figure 4b: Molecular pathways associated with normal ageing. Legends for the diamond nodes with a suffix of "-C/N" represent the peptide's cleavage site; i.e. "-C" for C-terminus and "-N" for the N-terminus."

Table 3. Abbreviations for protease and peptide names in figure 4a and 4b.

Protease	Peptides
MMP2 - 72 kDa type IV collagenase	COL1A2-N - Collagen alpha-2(I) chain
MMP14 - Matrix metalloproteinase-14	COL3A1-N - Collagen alpha-1(III) chain
MMP8 - Neutrophil collagenase	COL1A2-C - Collagen alpha-2(I) chain
ADAMTS5 - A disintegrin and metalloproteinase with thrombospondin motifs 5	COL3A1-C - Collagen alpha-1(III) chain
MMP9 - Matrix metalloproteinase-9	COL4A3-C - Collagen alpha-3(IV) chain
CTSK - Cathepsin K	COL1A1-N - Collagen alpha-1(I) chain
MMP7 - Matrilysin	COL25A1-N - Collagen alpha-1(XXV) chain
MMP13 - Collagenase 3	COL6A1-C - Collagen alpha-1(VI) chain
CTSL1 - Cathepsin L1	COL2A1-C - Collagen alpha-1(II) chain
CTSS - Cathepsin S	COL1A1-C - Collagen alpha-1(I) chain
MMP12 - Macrophage metalloelastase	COL9A3-N - Collagen alpha-3(IX) chain
ADAMTS4 - A disintegrin and metalloproteinase with thrombospondin motifs 4	COL4A1-C - Collagen alpha-1(IV) chain
PLC - 1-phosphatidylinositol 4,5-bisphosphate phosphodiesterase gamma-1	COL25A1-C - Collagen alpha-1(XXV) chain
F2 - Prothrombin	COL16A1-C - Collagen alpha-1(XVI) chain
MMP25 - Matrix metalloproteinase-25	
CAPN1 - Calpain-1 catalytic subunit	

5. Discussion

For an early diagnosis and successful targeted treatment, molecular characterization of individual disease phenotypes and prediction of novel biomarkers is essential. By the use of network biology approaches such as bioinformatics analysis tools, pathway databases, and statistical criteria, a comprehensive understanding of the complex molecular mechanisms in genetic disorders can be achieved. This would be advantageous in better prognosis and early clinical intervention of the individual phenotype. Based on this background, the aim of this thesis was to characterize muscle invasive bladder carcinoma on a molecular level by incorporating signatures from scientific literature screening and omics profiling. The characterization was achieved by integrating collected data to perform protein-protein interactions and pathway enrichment analysis.

In first part, of this study, automated data retrieval from the literature resulted in a first collection of molecular features associated with muscle invasive bladder cancer, and, combination with omics profiling data, allowed the creation of a mechanistic (pathway) map linked to muscle invasive bladder cancer. By deriving bladder cancer-associated protein coding genes on the basis of such pathway maps provides a systematic foundation for experimental analysis regarding association with development of muscle-invasive disease.

In the second part of the thesis, the performance of pathway enrichment was compared for significant pathway outputs yielded from ClueGO and ImPAla in the context to bladder cancer. Only two widely used and up-to-date pathway database resources, KEGG and Reactome were selected. Adding more pathway databases in the analysis would introduce higher redundancy in pathway outputs. In regard to manually updating database sources, the ClueGO application allows users to update individual pathway database source in order to obtain latest data whereas ImPAla is an omics-integration focusing towards metabolomics integration and pathway enrichment application that contains the latest update of January 2015[67]. In addition, ImPAla also allows the incorporation of differential expression information for molecules such as magnitude of differentially expressed fold changes and multiple comparison corrected p-values. Nevertheless, ImPAla does not offer the option to input separately up and down regulated genes and does not predict activation/deactivation of an affected pathway in contrast to ClueGO. In addition, ClueGO provides users to analyze different omics datasets such as genes, mRNAs, proteins, single nucleotide polymorphisms, metabolites and miRNAs. This gives the advantage in using one analysis and visualization tool for all high-throughput sequencing and profiling experiments. Furthermore, having a

single analysis tool also helps to prevent errors due to compatibility when transferring data between different software applications. Therefore, ClueGO is preferable to ImPAla for pathway enrichment and in the comprehensive characterization of molecular diseases.

In the last section of the thesis, the analysis of the urinary peptidome of ageing-associated peptides was detected. Differentially expressed age-associated peptides were identified using capillary electrophoresis coupled with mass spectrometry (CE-MS). The Proteasix software was then used in order to predict proteases that cleaved the identified urinary age associated peptides[68]. The generated data) were then subjected to systems biology and bioinformatics approaches such as pathway enrichment analysis in order to characterize molecular pathways that were associated with normal and pathological ageing. Findings demonstrated that with the help of appropriate peptidomics technologies, urine could be used as a powerful biological fluid in ageing research.

In conclusion, it is demonstrated in this thesis that by using bioinformatics and systems biology methodologies a better understanding of complex molecular mechanisms such as tumor invasion in bladder cancer is possible. In addition, ClueGO pathway enrichment tool has better performance than ImPAla in pathway enrichment analysis since the pathway output is less redundant and contains all the biologically significant information. The combination of a systems biology approach and individual proteins biochemical features offers a thorough molecular description of muscle invasive bladder cancer.

6. References

1. Grossman HB, Natale RB, Tangen CM, et al. Neoadjuvant chemotherapy plus cystectomy compared with cystectomy alone for locally advanced bladder cancer. *N Engl J Med*. 2003;349(9):859–66.
2. Kiss B, Skuginna V, Fleischmann A, et al. Bcl-2 predicts response to neoadjuvant chemotherapy and is overexpressed in lymph node metastases of urothelial cancer of the bladder. *Urol Oncol*. 2015;33(4):166.e1–8.
3. Ru Y, Dancik GM, Theodorescu D. Biomarkers for prognosis and treatment selection in advanced bladder cancer patients. *Curr Opin Urol*. 2011;21:420–427.
4. Henning A, Wehrberger M, Madersbacher S, et al. Do differences in clinical symptoms and referral patterns contribute to the gender gap in bladder cancer? *BJU Int*. 2013;112(1):68–73.
5. Raman JD, Shariat SF, Karakiewicz PI, et al. Does preoperative symptom classification impact prognosis in patients with clinically localized upper-tract urothelial carcinoma managed by radical nephroureterectomy? *Urol Oncol*. 2011;29(6):716–23.
6. Cauberg Evelyne CC, de la Rosette JJMCH, de Reijke TM. Emerging optical techniques in advanced cystoscopy for bladder cancer diagnosis: A review of the current literature. *Indian J Urol*. 2011;27(2):245–51.
7. Niegisch G, Lorch A, Droller MJ, Lavery HJ, Stensland KD, Albers P. Neoadjuvant chemotherapy in patients with muscle-invasive bladder cancer: which patients benefit? *Eur Urol*. 2013;64(3):355–7.
8. Sternberg CN, Apolo AB. Everything old is new again! Neoadjuvant chemotherapy in the treatment of muscle-invasive bladder cancer. *J Clin Oncol*. 2014;32(18):1868–70.
9. Knowles MA, Hurst CD. Molecular biology of bladder cancer: new insights into pathogenesis and clinical diversity. *Nat Rev Cancer*. 2014;15(1):25–41.
10. Anastasiadis A, de Reijke TM. Best practice in the treatment of nonmuscle invasive bladder cancer. *Ther Adv Urol*. 2012;4(1):13–32. doi:10.1177/1756287211431976.
11. Mitra AP, Cote RJ. Molecular pathogenesis and diagnostics of bladder cancer. *Annu Rev Pathol*. 2009;4:251–85.
12. Vishnu P, Mathew J, Tan WW. Current therapeutic strategies for invasive and metastatic bladder cancer. *Onco Targets Ther*. 2011;4:97–113.
13. Mitra AP, Datar RH, Cote RJ. Molecular staging of bladder cancer. *BJU Int*. 2005;96:7–12.
14. Choi W, Porten S, Kim S, et al. Identification of distinct basal and luminal subtypes of muscle-invasive bladder cancer with different sensitivities to frontline chemotherapy. *Cancer Cell*. 2014;25(2):152–65.
15. Lowe SW, Ruley HE, Jacks T, Housman DE. p53-dependent apoptosis modulates the cytotoxicity of anticancer agents. *Cell*. 1993;74(6):957–67.
16. McConkey DJ, Lee S, Choi W, et al. Molecular genetics of bladder cancer: Emerging mechanisms of tumor initiation and progression. *Urol Oncol*. 2010;28(4):429–40.
17. Bayraktar Z, Gurbuz G, Taşci AI, Sevin G. Staging error in the bladder tumor: the correlation between stage of TUR and cystectomy. *Int Urol Nephrol*. 2001;33(4):627–9.

18. Lerner SP, Skinner E, Skinner DG. Radical cystectomy in regionally advanced bladder cancer. *Urol Clin North Am*. 1992;19(4):713–23.
19. Kim PH, Cha EK, Sfakianos JP, et al. Genomic predictors of survival in patients with high-grade urothelial carcinoma of the bladder. *Eur Urol*. 2015;67(2):198–201.
20. Emmert-Streib F, Abogunrin F, de Matos Simoes R, et al. Collectives of diagnostic biomarkers identify high-risk subpopulations of hematuria patients: exploiting heterogeneity in large-scale biomarker data. *BMC Med*. 2013;11:12.
21. Mayer P, Mayer B, Mayer G. Systems biology: building a useful model from multiple markers and profiles. *Nephrol Dial Transplant*. 2012;27:3995–4002.
22. Kreeger PK, Lauffenburger DA. Cancer systems biology: a network modeling perspective. *Carcinogenesis*. 2010;31:2–8.
23. Chihara Y, Kanai Y, Fujimoto H, et al. Diagnostic markers of urothelial cancer based on DNA methylation analysis. *BMC Cancer*. 2013;13:275.
24. Weiss GJ, Liang WS, Demeure MJ, et al. A pilot study using next-generation sequencing in advanced cancers: feasibility and challenges. *PLoS One*. 2013;8(10):e76438.
25. Mischak H, Schanstra JP. CE-MS in biomarker discovery, validation, and clinical application. *Proteomics Clin Appl*. 2011;5:9–23.
26. Alberice JV, Amaral AFS, Armitage EG, et al. Searching for urine biomarkers of bladder cancer recurrence using a liquid chromatography-mass spectrometry and capillary electrophoresis-mass spectrometry metabolomics approach. *J Chromatogr A*. 2013;1318:163–70.
27. Pignot G, Cizeron-Clairac G, Vacher S, et al. microRNA expression profile in a large series of bladder tumors: Identification of a 3-miRNA signature associated with aggressiveness of muscle-invasive bladder cancer. *Int J Cancer*. 2013;132(11):2479–2491.
28. Liu Y, Noon AP, Aguiar Cabeza E, et al. Next-generation RNA sequencing of archival formalin-fixed paraffin-embedded urothelial bladder cancer. *Eur Urol*. 2014;66(6):982–6.
29. Hanash S. Disease proteomics. *Nature*. 2003;422(6928):226–32.
30. Rustici G, Kolesnikov N, Brandizi M, et al. ArrayExpress update--trends in database growth and links to data analysis tools. *Nucleic Acids Res*. 2013;41(Database issue):D987–90.
31. Barrett T, Wilhite SE, Ledoux P, et al. NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Res*. 2013;41(Database issue):D991–5.
32. Muthusamy B, Thomas JK, Prasad TSK, Pandey A. Access guide to human proteinpedia. *Curr Protoc Bioinformatics*. 2013;Chapter 1:Unit 1.21.
33. Prasad TSK, Kandasamy K, Pandey A. Human Protein Reference Database and Human Proteinpedia as discovery tools for systems biology. *Methods Mol Biol*. 2009;577:67–79.
34. Stelzer G, Dalah I, Stein TI, et al. In-silico human genomics with GeneCards. *Hum Genomics*. 2011;5(6):709–17.
35. Rhodes DR, Yu J, Shanker K, et al. ONCOMINE: a cancer microarray database and integrated data-mining platform. *Neoplasia*. 2004;6(1):1–6.
36. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian

Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res*. 2005;33:D514–D517.
37. Weinstein JN, Akbani R, Broom BM, et al. Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature*. 2014;507(7492):315–322.
38. Zabolotneva AA, Zhavoronkov AA, Shegay P V, et al. A systematic experimental evaluation of microRNA markers of human bladder cancer. *Front Genet*. 2013;4:247.
39. Zabolotneva AA, Zhavoronkov A, Garazha A V, Roumiantsev SA, Buzdin AA. Characteristic patterns of microRNA expression in human bladder cancer. *Front Genet*. 2012;3:310.
40. Dancik GM. An online tool for evaluating diagnostic and prognostic gene expression biomarkers in bladder cancer. *BMC Urol*. 2015;15(1):59.
41. Mayer G, Heinze G, Mischak H, et al. Omics-bioinformatics in the context of clinical data. *Methods Mol Biol*. 2011;719:479–497.
42. Abu-Asab MS, Chaouchi M, Alesci S, et al. Biomarkers in the age of omics: time for a systems biology approach. *OMICS*. 2011;15:105–112.
43. De Las Rivas J, Fontanillo C. Protein-protein interactions essentials: key concepts to building and analyzing interactome networks. *PLoS Comput Biol*. 2010;6:e1000807.
44. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for deciphering the genome. *Nucleic Acids Res*. 2004;32(Database issue):D277–80.
45. Mi H, Muruganujan A, Thomas PD. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Res*. 2013;41(Database issue):D377–86.
46. Croft D, O'Kelly G, Wu G, et al. Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res*. 2011;39(Database issue):D691–7.
47. Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S. AmiGO: online access to ontology and annotation data. *Bioinformatics*. 2009;25:288–289.
48. Frantzi M, Bhat A, Latosinska A. Clinical proteomic biomarkers: relevant issues on study design & technical considerations in biomarker development. *Clin Transl Med*. 2014;3(1):7.
49. Krishnamurthy L, Nadeau J, Ozsoyoglu G, et al. Pathways database system: an integrated system for biological pathways. *Bioinformatics*. 2003;19(8):930–7.
50. Stein LD. Integrating biological databases. *Nat Rev Genet*. 2003;4(5):337–45.
51. Philippi S, Köhler J. Addressing the problems with life-science databases for traditional uses and systems biology. *Nat Rev Genet*. 2006;7(6):482–8.
52. Forbes SA, Bindal N, Bamford S, et al. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res*. 2011;39(Database issue):D945–50.
53. Gui Y, Guo G, Huang Y, et al. Frequent mutations of chromatin remodeling genes in transitional cell carcinoma of the bladder. *Nat Genet*. 2011;43(9):875–8.
54. Hurst CD, Platt FM, Taylor CF, Knowles MA. Novel tumor subgroups of urothelial carcinoma of the bladder defined by integrated genomic analysis. *Clin Cancer Res*. 2012;18(21):5865–77.
55. Lindgren D, Sjödahl G, Lauss M, et al. Integrated genomic and gene expression profiling identifies two major genomic circuits in urothelial carcinoma. *PLoS One*. 2012;7(6):e38863.

56. Williams S V, Hurst CD, Knowles MA. Oncogenic FGFR3 gene fusions in bladder cancer. *Hum Mol Genet*. 2013;22(4):795–803.
57. Kerrien S, Aranda B, Breuza L, et al. The IntAct molecular interaction database in 2012. *Nucleic Acids Res*. 2012;40(Database issue):D841–6.
58. Stark C, Breitkreutz B-J, Chatr-Aryamontri A, et al. The BioGRID Interaction Database: 2011 update. *Nucleic Acids Res*. 2011;39:D698–D704.
59. Franceschini A, Szklarczyk D, Frankild S, et al. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res*. 2013;41(Database issue):D808–15.
60. Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*. 2003;13(11):2498–504.
61. Saito R, Smoot ME, Ono K, et al. A travel guide to Cytoscape plugins. *Nat Methods*. 2012;9(11):1069–76.
62. Bindea G, Mlecnik B, Hackl H, et al. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics*. 2009;25(8):1091–3.
63. Bindea G, Galon J, Mlecnik B. CluePedia Cytoscape plugin: pathway insights using integrated experimental and in silico data. *Bioinformatics*. 2013;29(5):661–3.
64. Calabrese V, Dattilo S, Petralia A, et al. Analytical approaches to the diagnosis and treatment of aging and aging-related disease: redox status and proteomics. *Free Radic Res*. 2015;49(5):511–24.
65. Nkuipou-Kenfack E, Koeck T, Mischak H, et al. Proteome analysis in the assessment of ageing. *Ageing Res Rev*. 2014;18:74–85. doi:10.1016/j.arr.2014.09.002.
66. Zürbig P, Decramer S, Dakna M, et al. The human urinary proteome reveals high similarity between kidney aging and chronic kidney disease. *Proteomics*. 2009;9(8):2108–17.
67. Kamburov A, Cavill R, Ebbels TMD, Herwig R, Keun HC. Integrated pathway-level analysis of transcriptomics and metabolomics data with IMPaLA. *Bioinformatics*. 2011;27(20):2917–8.
68. Klein J, Eales J, Zürbig P, Vlahou A, Mischak H, Stevens R. Proteasix: A tool for automated and large-scale prediction of proteases involved in naturally occurring peptide generation. *Proteomics*. 2013;13(7):1077–82.

7. Affidavit

I, Akshay Bhat certify under penalty of perjury by my own signature that I have submitted the thesis on the topic "Bioinformatics modeling of proteomics changes in muscle invasive bladder cancer". I wrote this thesis independently and without assistance from third parties, I used no other aids than the listed sources and resources.

All points based literally or in spirit on publications or presentations of other authors are, as such, in proper citations (see "uniform requirements for manuscripts (URM)" the ICMJE www.icmje.org) indicated. The sections on methodology (in particular practical work, laboratory requirements, statistical processing) and results (in particular images, graphics and tables) correspond to the URM (s.o) and are answered by me. My contributions in the selected publications for this dissertation correspond to those that are specified in the following joint declaration with the responsible person and supervisor. All publications resulting from this thesis and which I am author of correspond to the URM (see above) and I am solely responsible.

The importance of this affidavit and the criminal consequences of a false affidavit (section 156,161 of the Criminal Code) are known to me and I understand the rights and responsibilities stated therein.

Date

_____

Signature

Declaration for the contribution in publications:

Akshay Bhat had the following contribution in the following publications.

Publication 1: Akshay Bhat, Andreas Heinzel, Bernd Mayer, Paul Perco, Irmgard Mühlberger, Holger Husi, Axel S. Merseburger, Joost P Schanstra, Jerome Zoidakis, Antonia Vlahou, Harald Mischak, *Vera Jankowski*. Protein Interactome for Muscle Invasive Bladder Cancer. PLoS One. 2015.

Contribution in detail: data mining and processing / omics integration / protein-protein interaction analysis / statistical analysis / pathway enrichment analysis / drafting the publication.

Percentage of contribution: 15%

Publication 2: Akshay Bhat, *Vera Jankowski*, Antonia Vlahou, Harald Mischak, Jerome Zoidakis. Comparison of ClueGO and ImPAla for integrated pathway enrichment analysis Accepted in Jacobs Journal of Bioinformatics and Proteomics –[Epub ahead of print]

Contribution in detail: the study protocol / data mining and processing / systems biology analysis / statistical analysis / software comparison / database integration / drafting the publication.

Percentage of contribution: 17%

Publication 3: Esther Nkuipou-Kenfack, **Akshay Bhat**, *Vera Jankowski*, Julie Klein, William Mullen, Antonia Vlahou, Thomas Koeck, Mohammed Dakna, Joost P. Schanstra, Petra Zürbig, Andreas Pich, Harald Mischak. Identification of urinary age-specific peptides in a healthy population. DOI: 10.18632/oncotarget.5896

Contribution in detail: systems biology analysis / critical proofreading

Percentage of contribution: 7%

Akshay Bhat (MSc.) PD. Dr. Vera Jankowski

List of selected Publications:

Publication 1:

Protein Interactome for Muscle Invasive Bladder Cancer.

Akshay Bhat, Andreas Heinzel, Bernd Mayer, Paul Perco, Irmgard Mühlberger, Holger Husi, Axel S. Merseburger, Joost P Schanstra, Jerome Zoidakis, Antonia Vlahou, Harald Mischak, *Vera Jankowski.*

PLoS One. 2015 Jan 8;10(1):e0116404. doi: 10.1371/journal.pone.0116404. eCollection 2015.

Publication 2:

Comparison of ClueGO and ImPAla for integrated pathway enrichment analysis

Akshay Bhat, *Vera Jankowski*, Antonia Vlahou, Harald Mischak, Jerome Zoidakis.

Zoidakis J. Comparison of Cluego and Impala for Integrated Pathway Enrichment Analysis. J J Bioinform Proteom. 2016, 1(1): 002.

Publication 3:

Identification of urinary age-specific peptides in a healthy population.

Esther Nkuipou-Kenfack, **Akshay Bhat**, *Vera Jankowski*, Julie Klein, William Mullen, Antonia Vlahou, Thomas Koeck, Mohammed Dakna, Joost P. Schanstra, Petra Zürbig, Andreas Pich, Harald Mischak.

DOI: 10.18632/oncotarget.5896; Oncotarget. 2015 Sep 29.

PLOS | ONE

RESEARCH ARTICLE

Protein Interactome of Muscle Invasive Bladder Cancer

Akshay Bhat[1,2], Andreas Heinzel[3], Bernd Mayer[3], Paul Perco[3], Irmgard Mühlberger[3], Holger Husi[4], Axel S. Merseburger[5], Jerome Zoidakis[6], Antonia Vlahou[6], Joost P. Schanstra[7,8], Harald Mischak[2,4], Vera Jankowski[9]*

1 Charité-Universitätsmedizin Berlin, Med. Klinik IV, Berlin, Germany, **2** Mosaiques diagnostics GmbH, Hannover, Germany, **3** emergentec biodevelopment GmbH, Vienna, Austria, **4** BHF Glasgow Cardiovascular Research Centre, University of Glasgow, Glasgow, United Kingdom, **5** Department of Urology and Urological Oncology, Hannover Medical School, Hannover, Germany, **6** Biomedical Research Foundation Academy of Athens, Biotechnology Division, Athens, Greece, **7** Institut National de la Santé et de la Recherche Médicale (INSERM), U1048, Institute of Cardiovascular and Metabolic Diseases, Toulouse, France, **8** Université de Toulouse III Paul Sabatier, Toulouse, France, **9** Institute for Molecular Cardiovascular Research (IMCAR), Aachen, Germany

* vjankowski@ukaachen.de

OPEN ACCESS

Citation: Bhat A, Heinzel A, Mayer B, Perco P, Mühlberger I, Husi H, et al. (2015) Protein Interactome of Muscle Invasive Bladder Cancer. PLoS ONE 10(1): e0116404. doi:10.1371/journal.pone.0116404

Academic Editor: Francisco X. Real, Centro Nacional de Investigaciones Oncológicas (CNIO), SPAIN

Received: May 15, 2014

Accepted: December 9, 2014

Published: January 8, 2015

Data Availability Statement: Data are from the multiple literature articles study whose authors may be contacted through the PubMed ids provided in this manuscript.

Funding: This work was supported by Marie Curie Actions — BCMolMed under grant agreement no. FP7-PEOPLE-2012-ITN-EID and the European Community's Seventh Framework Programme under grant agreement no. 306157. Harald Mischak is the founder and co-owner of Mosaiques Diagnostics, who developed the CE-MS technology for clinical application. Akshay Bhat is an employee of Mosaiques Diagnostics. Bernd Mayer is the managing partner of emergentec biodevelopment GmbH, Austria. Paul

Abstract

Muscle invasive bladder carcinoma is a complex, multifactorial disease caused by disruptions and alterations of several molecular pathways that result in heterogeneous phenotypes and variable disease outcome. Combining this disparate knowledge may offer insights for deciphering relevant molecular processes regarding targeted therapeutic approaches guided by molecular signatures allowing improved phenotype profiling. The aim of the study is to characterize muscle invasive bladder carcinoma on a molecular level by incorporating scientific literature screening and signatures from omics profiling. Public domain omics signatures together with molecular features associated with muscle invasive bladder cancer were derived from literature mining to provide 286 unique protein-coding genes. These were integrated in a protein-interaction network to obtain a molecular functional map of the phenotype. This feature map educated on three novel disease-associated pathways with plausible involvement in bladder cancer, namely Regulation of actin cytoskeleton, Neurotrophin signalling pathway and Endocytosis. Systematic integration approaches allow to study the molecular context of individual features reported as associated with a clinical phenotype and could potentially help to improve the molecular mechanistic description of the disorder.

Introduction

Bladder cancer (BC) presents with an estimate of 72,570 new cases diagnosed and 15,210 deaths across the United States [1] in the year 2013, clearly demonstrating a need for improved diagnosis and therapy. Bladder cancer is the ninth most frequent malignancy with an approximate ratio of 5:1 with respect to non-muscle invasive versus muscle invasive phenotypes [2].

Perco, Andreas Heinzel and Irmgard Mühlberger are employees of emergentec. Neither Mosaiques Diagnostics GmbH nor emergentec biodevelopment GmbH were involved in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The companies involved only provided financial support in the form of authors' salaries and/or research materials. The specific roles of these authors are articulated in the 'author contributions' section.

Competing Interests: The authors have declared the following potential conflict of interest: Harald Mischak is the founder and co-owner of Mosaiques Diagnostics, who developed the CE-MS technology for clinical application. Akshay. Bhat is an employee of Mosaiques Diagnostics. Bernd Mayer is the managing partner of emergentec biodevelopment GmbH, Austria. Paul Perco, Andreas Heinzel and Irmgard Mühlberger are employees of emergentec. These issues do not alter the authors' adherence to all the PLOS ONE policies on sharing data and materials.

Major confounders are smoking and other occupational exposures along with genetic predispositions, such as e.g. N-acetyltransferase 1 (NAT1), N-acetyltransferase 2 (NAT2) and glutathione S-transferase μ1 (GSTM1) polymorphisms [3]. Though variable for bladder cancer patients, initial symptoms include haematuria and flank pain, commonly represented during advanced cancer stages caused by ureteric obstructions due to invasion of the bladder muscular wall or ureter, together with recurrent urinary tract infections [4, 5]. Evidence suggests that malignant transformation of the bladder is multifactorial and a multitude of genes are involved in the development of muscle invasive or non-muscle invasive phenotype [6, 7]. The major histological type is transitional cell carcinoma occurring in approximately 90% of diagnosed bladder tumours (with the rest being mainly squamous cell carcinomas and adenocarcinomas), with categories of non-invasive papillary (Ta) or flat (Tis), subepithelial invasive (T1), muscle invasive (T2–T4) and metastatic (N+, M+) diseases, all differing in biology, progression characteristics and hence clinical management. Majority of the cases are non-muscle invasive (Tis, Ta, T1) and 10–15% are muscle-invasive tumours (T2–T4), with the latter associated with fast recurrence and poorer prognosis based on progressing towards metastasis formation.

Cystoscopy is the gold standard with a reported sensitivity and specificity in the range of 62–84% and 43–98%, respectively [8]. Due to the invasive nature of the procedure, but also for adding accuracy in the detection, biomarkers assessed in blood or urine are considered as beneficial for supporting clinical assessment [9]. This is also relevant for disease prognosis as biomarkers measured at the DNA, RNA and/or protein levels provide the potential to choose best surveillance measures and treatment regimens for specific patient populations regarding halting the development of muscle invasive disease [10]. Treatment of papillary and non-muscle invasive high-grade carcinoma involves endoscopic transurethral resection of visible tumours followed by adjuvant treatment with intravesical instillation therapy (Mitomycin/Epirubicin or Bacillus Calmette-Guerin (BCG)) depending on the estimated risk for progression. Irrespective of aggressive treatment and vigorous follow-up, 70% of these tumours recur, and 25% of high-grade non-muscle invasive cancers progress into invasive phenotypes [2, 11].

The comparison of the genetic characteristics of muscle-invasive and non-invasive tumours revealed that non-invasive tumours over-express HRAS and FGFR3 or produce highly activated forms of these proteins. As a result, the Ras/MAPK pathways are up-regulated in non-invasive tumours [12]. Muscle-invasive BC is associated with alterations of p53, retinoblastoma protein (RB1) and tumour suppressors controlling cell cycle processes, in addition to elevated expressions in epidermal growth factor receptor (EGFR), human epidermal growth factor receptor 2 (HER2/ErbB2), matrix metallopeptidase 2 (MMP2) and MMP9 and deletions in p16Ink4a and P15Ink4b [3].

High-throughput experimental platform technologies ranging from genomic sequencing to proteomic and metabolomic profiling are now being used for molecular characterization of clinical phenotypes [13–19]. A variety of datasets have become available e.g. in Array Express/ Gene Expression Omnibus (GEO) for transcriptomics, Human Proteinpedia for proteomics, or in large data consolidation platforms such as GeneCards [20]. In regard to disease specific omics data, valuable general sources in oncology include TCGA (http://cancergenome.nih.gov/), Oncomine [21], and OMIM [22]. Though omics profiling has provided an abundance of data, technical boundaries involving incompleteness of the individual molecular catalogues together with the static representation of cellular activity limits the insights on molecular processes and their interaction dynamics [23–25]. Despite these challenges, omics-based profiling has significantly advanced bladder cancer research, providing the basis for an integrative analysis approach in delineating a more comprehensive overview of molecular processes and pathways that characterize variations of muscle-invasive urothelial carcinoma [12].

On the effector level, proteins interact and co-operatively form specific molecular processes and pathways. Intermolecular interactions include various types being represented as networks (graphs) with molecular features denoted as nodes (vertices) together with their interactions (edges). A large number of biological pathway resources has become available, including KEGG [26], PANTHER [27], REACTOME [28] and AmiGO [29] described in PathGuide (http://www.pathguide.org/), all displaying well-defined human molecular metabolic and signalling pathways together with disease-specific pathways (e.g. pathways in cancer). Molecular features being identified as associated with bladder cancer can be interpreted on the level of such pathways, adding to a functional interpretation of molecular feature sets characterizing the phenotype.

To add to our understanding of muscle-invasive bladder carcinoma (MIBC), we derived a phenotype-specific network model (interactome) by integrating omics signatures characterizing MIBC, reported in scientific literature and databases. Our procedure incorporated scientific literature screening and signatures from omics profiling, resulting in 1,054 protein-coding genes being associated with MIBC, further consolidating to 286 genes on the interactome level. The results display deriving a systems-level model for molecular phenotyping of bladder cancer muscle invasion, presented as multiple affected pathways.

Materials and Methods

Data sources for characterizing bladder cancer pathophysiology

For consolidating molecular features associated with muscle invasive bladder cancer, NCBI PubMed, Web of Science, Google Scholar and the omics repositories Gene Expression Omnibus (GEO) [30] and ArrayExpress [31] were queried. The keywords for the literature search included "bladder OR urothelial OR transitional cell" AND "neoplasm OR tumor OR carcinoma" AND "muscle" AND "invas∗ OR aggress∗ OR progress∗ OR inflammation" (Database version of April, 2014). By construction this search query focused specifically on muscle invasive bladder neoplasm. For extracting protein-coding genes associated with these publications gene-2-pubmed as provided by NCBI was used [32]. The list of publications relevant to bladder cancer muscle invasion was isolated from the complete list of papers indexed in PubMed along with the associated gene IDs (ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2pubmed.gz). Profiling experiments were further screened for adequacy in sample size (at least 50 samples included in study design), magnitude of differential abundance (>2-fold change) and the specific phenotypic conditions; T1, $T2_{a/b}$, $T3_{a/b}$, $T4_{a/b}$ (Figs. 1 and 2). In addition, only papers mentioning the keywords "molecular" and "biomarker" were retained for deriving the literature mined MIBC molecules and pathways.

Interaction data and induced subgraph

Protein interaction information was obtained by querying IntAct [33], BioGRID [34], and Reactome [28] leading to a total of 233,794 interactions covering 13,907 protein-coding genes within the human interactome (Databases in version of April, 2014). Mapping the MIBC associated molecular features on this consolidated interaction network [13] provided an MIBC-specific induced subgraph. MIBC associated features not connected to at least another such feature were disregarded from further analysis.

Functional analysis

Cytoscape's plug-ins ClueGO and CluePedia was used to identify pathways that are being over-represented in the set of features located in the induced subgraph [35, 36]. KEGG pathway

terms served as the clustering criterion using a two-sided hypergeometry test followed by Bonferroni correction (significance level of 0.05) for identifying significantly affected pathways. General disease pathways (such as pathways in cancer, miRNA's in cancer, bladder cancer etc.) were discarded to obtain a set of generic pathway terms [13].

Protein coding gene selection based on literature mining

From the set of MIBC-associated protein-coding genes, each gene symbol was evaluated for being a member of the MIBC pathway set. The evidence of identified pathways and extracted genes involved in MIBC was assessed based on the level of annotation depth, defined as the number of individual studies identifying such protein-coding genes as involved in MIBC. Specifically, such evidence was derived from metadata available in PubMed. Gene-2-pubmed was used for linking the molecules contained in the induced subgraph to publications relevant to bladder cancer muscle invasion. The quality of publications obtained for each molecule was assessed based on manual reviewing. Only papers where a direct link of the molecule to bladder cancer muscle invasion was proven were retained. For the entire pathway set, the ratio between the number of molecules being linked to at least one urinary bladder neoplasm publication and the number of features in the pathway was computed and used for relevance ranking. For individual protein-coding genes identified in literature the number of linked urinary bladder neoplasm publications was used as relevance ranking criterion.

Results

Data Mining

Mining of published articles and omics repositories led to a collection of 285 references after manual screening (Fig. 1). This screening was performed to discard duplicated studies retrieved

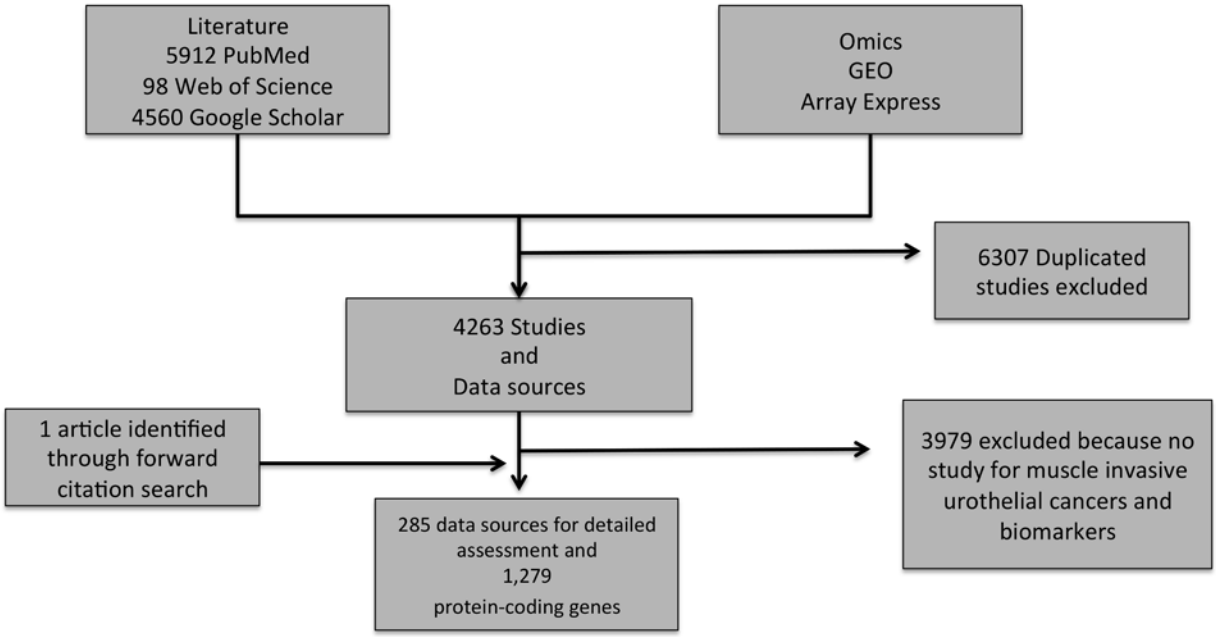

Figure 1. Data assembly workflow. PubMed, Google Scholar and Web of Science literature analysis and Omics data source screening with focus on transcriptomics. From the 4263 abstracts screened 3979 articles were excluded not specifically focusing on muscle-invasive bladder cancer phenotype (stages T2–T4). 188 studies out of 285 articles were discarded, as these did not meet required study designs and 2-fold change in magnitude of differential abundance of identified features. This restriction resulted in 1,279 protein-coding genes and was further used in the systems based analysis for MIBC.

doi:10.1371/journal.pone.0116404.g001

from the varying repositories as well as articles not explicitly focusing on muscle-invasive bladder carcinoma. All molecular features were converted to their official gene symbol by using the UniProt ID Mapping service [37]. The resulting set of references yielded in total 1,279 proteins of which 1,054 were unique protein-coding genes associated with MIBC (S1 Table). For collecting specifically proteins involved in MIBC, we further screened these 285 articles with the keywords ("molecular and biomarker") to retrieve 122 proteins that had a tag "biomarker" mentioned in these articles (S2 Table). This restriction helped in discarding general articles containing gene symbols that were not associated to the muscle-invasive phenotype. The same set of 285 articles was used to collect all pathways connected to bladder carcinoma. Thus, 11 pathways reported in the literature to be associated with bladder cancer were obtained (S3 Table).

The largest number of features associated with MIBC resulted from transcriptomics with a total of 716 gene symbols. Metabolites were mapped to protein coding genes using the Human Metabolome Database (HMDB) [38] and provided 329 gene symbols. The miRNAs from the transcriptomics studies were mapped to their respective gene targets using the service from miRbase [39, 40]. In addition, DNA-methylation studies provided 77 gene symbols respectively. DNA-mutation studies reported 35 gene symbols. Scientific literature analysis provided 178 protein-coding genes, of which 122 were further annotated as indicators of muscle-invasive bladder cancer. The detailed information on all differentially expressed molecular features is available in S2 and S4–S8 Tables.

Data Source Overlap

With respect to the feature set overlap, 52 gene symbols were identified in both, literature mining and individual omics signatures (S1 Table). Of the 1,054 unique gene symbols, 716 protein-coding genes were from transcriptomics studies, 13 of these were found on DNA-methylation level, 13 on DNA-mutation level, and 19 on the metabolome level (Fig. 2a). This relatively weak overlap on the level of individual features, however, is a frequent finding in cross-Omics data consolidation, in part stemming from constraints of applied methods, and different sample matrices under investigation in each case [41, 42].

Induced Subgraph

To increase evidence in regard to the association of molecular features with MIBC we included protein-interaction information as a filter mechanism, i.e. combining evidence from statistical analysis with biological (interaction) data. For this analysis, the molecular feature set was reduced from 1,054 unique protein coding genes to 592 gene symbols. The reduction in the protein list was mainly caused because the protein coding genes indirectly linked from metabolomics and miRNA profiling were not incorporated in the pathway analysis due to low evidence linking to respective targets and enzymes. Further disregarding features not showing interactions to other members of the MIBC set resulted in 286 protein-coding genes represented on the muscle invasive bladder cancer-specific subgraph. The list of the initial 1,054 proteins, the 592 proteins disregarding metabolomics data and miRNA screens, and the list of 286 proteins that form protein-protein interactions in the MIBC subgraph is available in S9 Table. This set of protein-coding genes with strong evidence regarding association with MIBC and holding interactions to other such features was included in pathway analysis (Fig. 2b).

Pathway enrichment

KEGG pathway enrichment analysis of the MIBC molecular feature set represented on the induced subgraph utilizing ClueGO and CluePedia resulted in 15 molecular pathways being

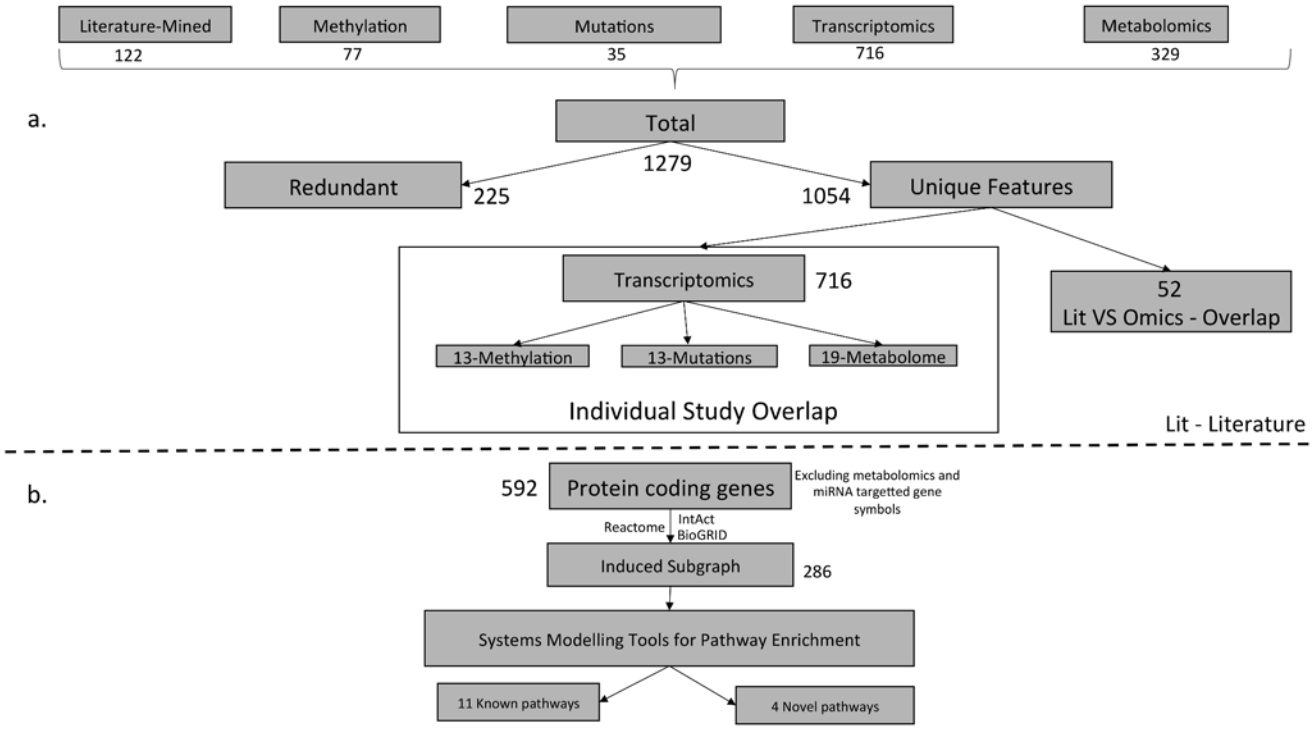

Figure 2. Feature set Overlap. A. Redundant features were discarded from 1,279 protein coding genes resulting in 1,054 unique features.The overlap between individual omics studies and literature were calculated. B. The 1,054 protein coding genes were further reduced to 592 by discarding enzymes linked to metabolites as well as miRNA targeted gene symbols, further included for deriving the induced MIBC subgraph resting on BioGRID, IntAct and Reactome protein interaction information.

doi:10.1371/journal.pone.0116404.g002

significantly affected in the context of muscle invasive bladder carcinoma (Fig. 3, S3 Table). In detail, 11 of these pathways were previously identified in literature, in addition to 4 presumably novel pathways that resulted from the interactome analysis.

The network in Fig. 3 represents each pathway as individual node, while the edges between pathways denote an approximation of biological interaction between the pathways based on the cross-pathway feature overlap. This pathway map allowed evaluating the functional context of the 122 literature-mined protein candidates in the context of MIBC.

Fig. 3 describes pathway terms enriched using the MIBC-specific induced subgraph. Categorizing the pathway terms in known and novel pathways according to literature, we obtained 11 pathways that were reported in the literature, namely Focal adhesion consisting of 40 protein coding genes, MAPK signalling pathway with 26, ECM-receptor interaction and Cell cycle with 17 features each, p53 with 16, Tight junction and Adherens junction with 15 features each, Leukocyte transendothelial migration with 12, VEGF signalling pathway with 11, and Gap junction containing 10 protein coding genes (see S3 Table). The novel set of pathways that were enriched in the analysis contained 4 pathway terms of which 3 pathways were resting on significant association with the muscle-invasive bladder cancer phenotype, namely Regulation of actin cytoskeleton holding 18 protein coding genes, Endocytosis with 16 and Neurotrophin signalling with 13 (Table 1). The highest overlap in gene symbols was found between regulation of actin cytoskeleton pathway and serotonergic synapse containing ARAF, HRAS, RAF1 and MAPK1, neurotrophin signalling pathway and regulation of actin cytoskeleton pathway, containing MAPK1, RAF1 and HRAS. The overlap of gene symbols between endocytosis and regulation of actin cytoskeleton pathway was FGFR3, EGFR and HRAS, while those between

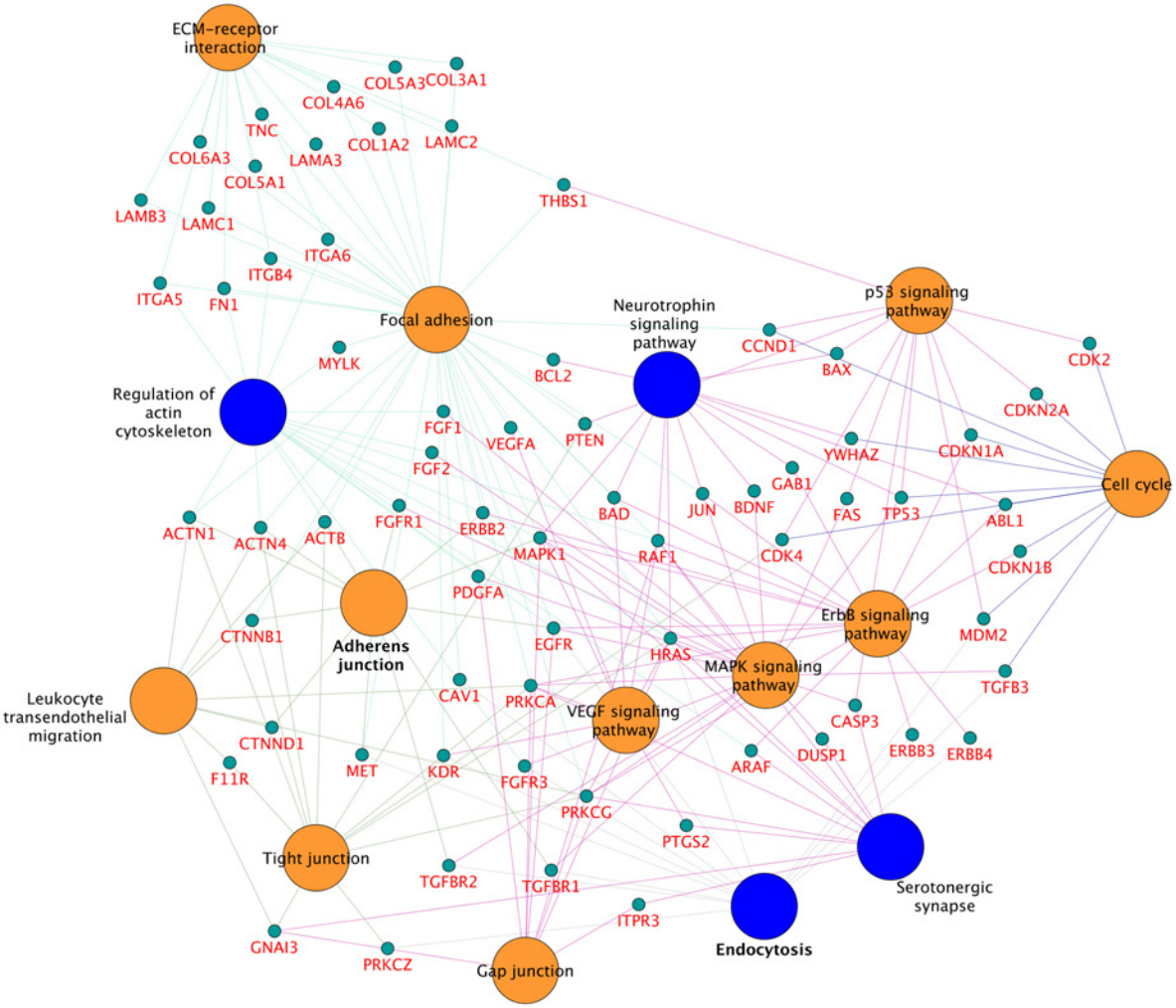

Figure 3. Muscle Invasive Bladder carcinoma interactome, set of 286 protein coding genes. Nodes in orange denote pathways identified as relevant in both literature and enrichment analysis, nodes in blue depicts pathways of relevance according to enrichment analysis. Node size scales with the number of gene symbols encoded in each pathway term.

doi:10.1371/journal.pone.0116404.g003

Table 1. KEGG pathways significantly associated with muscle invasive bladder carcinoma utilizing the gene set embedded in the induced subgraph.

KEGG Pathway Name	Number of MIBC features	Bonferroni corrected p-value	Overlapping protein-coding genes
Regulation of actin cytoskeleton	18	0.005874	**PDGFA, FGF1, RAF1, EGFR, ACTN4, FGFR1, ITGB4, FGFR3, MYLK, HRAS, ACTN1, FGF2, ITGA5, ARAF, FN1, MAPK1, ACTB, ITGA6**
Endocytosis	16	0.0344	**EGFR, MDM2, TGFBR2, FGFR3, HRAS, ERBB3, TGFB3, TGFBR1, ERBB4, CAV1, MET, PRKCZ, KDR**
Neurotrophin signalling pathway	13	0.01022	**BDNF, RAF1, BAD, HRAS, ABL1, GAB1, BCL2, TP53, BAX, YWHAZ, JUN, MAPK1**
Serotonergic synapse	12	0.0278	**RAF1, GNAI3, PRKCG, ITPR3, HRAS, CASP3, PTGS2, DUSP1, ARAF, MAPK1, PRKCA**

Pathway terms, total number of MIBC features associated with the term, Bonferroni corrected p-value and specific gene symbols found overlapping amongst the 4 pathway terms.

doi:10.1371/journal.pone.0116404.t001

neurotrophin signalling pathway and serotonergic synapse were HRAS, RAF1 and MAPK1. The least overlap of gene symbols between pathways was seen for neurotrophin signalling pathway and endocytosis, only sharing HRAS. Subsequently, there was no protein-coding gene overlapping for endocytosis and serotonergic synapse.

We performed an additional pathway enrichment analysis that involved an alternative set of gene symbols. From the full set of protein coding genes (707 molecules excluding metabolite and miRNA targetted gene symbols), we restricted to members being present in more than one study type (e.g. ERBB2 was found in proteomics, mRNA and literature mining). This restriction resulted in 72 gene symbols, again forwarded to pathway enrichment analysis. Fig. 4 details pathway terms enriched using this reduced set of protein coding genes. Seven pathway terms were enriched in this analysis. Categorizing these pathway terms into known from literature and novel pathways, 5 pathway terms were reported in literature and 2 pathways were novel findings. The 5 previously known pathway terms included Focal adhesion, Cell cycle, and p53 signalling pathway, ECM-receptor interaction, and ErbB signalling (S10 Table). In regard to the 2 novel pathways that were enriched from this analysis, the pathway terms were T cell receptor signaling pathway and GnRH signaling. Table 2 details all pathways with their overlapping gene symbols that were retrieved from this analysis. Regarding pathways with literature evidence the enrichment resting on the 72 gene symbols had a full overlap with the

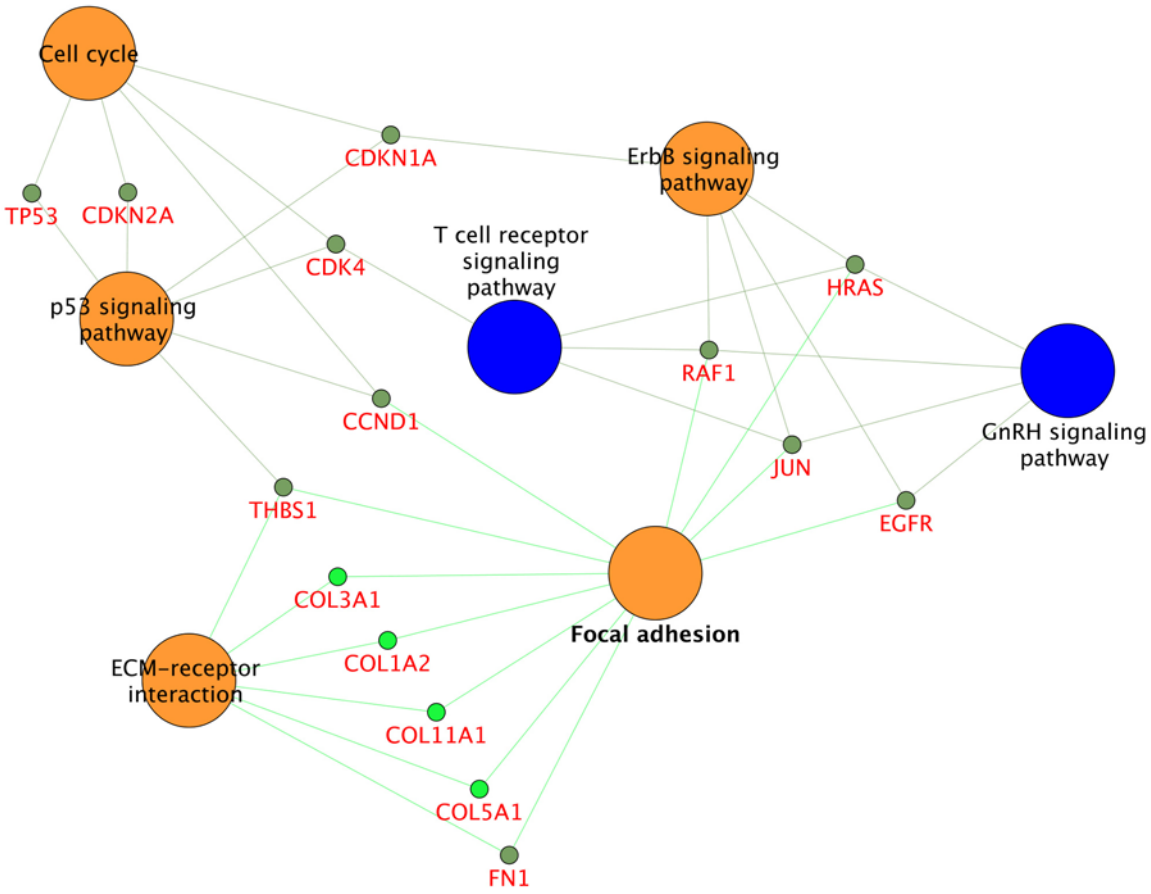

Figure 4. Muscle Invasive Bladder carcinoma pathway enrichment, set of 72 protein coding genes. Nodes in orange denote pathways identified as relevant in both literature and enrichment analysis; nodes in blue depict pathways of relevance according to enrichment analysis. The size of each node size scales with the number of gene symbols encoded in each pathway term.

doi:10.1371/journal.pone.0116404.g004

Table 2. KEGG pathways significantly associated with MIBC according to gene symbols found in more than one omics study type.

KEGG Pathway Name	Number of features	Bonferroni corrected p-value	Overlapping protein-coding genes
Focal adhesion	16	2.31E-011	**COL3A1, HRAS, CCND1, COL11A1, FN1, THBS1, JUN, COL5A1, RAF1, EGFR, COL1A2**
Cell cycle	7	0.00141	**CCND1, TP53, CDK4, CDKN2A, CDKN1A**
p53 signaling pathway	6	4.17E-04	**CCND1, THBS1, TP53, CDK4, CDKN2A, CDKN1A**
ErbB signalling pathway	6	0.00172	**HRAS, JUN, RAF1, CDKN1A, EGFR**
ECM-receptor interaction	6	0.00151	**COL3A1, COL11A1, FN1, THBS1, COL5A1, COL1A2**
GnRH signalling pathway	5	0.0360	**HRAS, JUN, RAF1, EGFR**
T cell receptor signalling pathway	5	0.048	**HRAS, JUN, CDK4, RAF1**

Pathway terms, number of molecules associated with the term, Bonferroni corrected p-value and specific gene symbols found overlapping amongst the 7 pathway terms.

doi:10.1371/journal.pone.0116404.t002

15 pathway terms retrieved using the induced subgraph MIBC feature set. The 2 pathways not showing an overlap were the novel pathways resting on the second gene set namely T cell receptor signalling and GnRH signalling (Fig. 4, S3 and S10 Tables).

Discussion

Understanding the molecular pathophysiology of muscle-invasive bladder carcinoma and revealing the network of pathways involved in muscle invasion could lead to targeted therapy. In addition, addressing specific dys-regulated pathways linked to progressive disease holds the promise of supporting an improved, biomarker-based risk assessment followed by stratified clinical intervention [2]. High throughput screening platforms have provided a wealth of information in describing the molecular status reflecting a clinical phenotype, including bladder carcinoma [43, 44]. Experiments based on expression profiling using microarrays, and fractionation techniques coupled to mass spectrometry utilizing tissue and urine as sample matrix have supported molecular pathway-based discovery in bladder muscle invasive neoplasms [12, 45]. The present study intended to characterize muscle invasive bladder carcinoma by incorporating scientific literature screening and signatures from omics profiling further linked in an interaction context, resulting in a set of 286 protein-coding genes. Such analysis on the level of networks and pathways was chosen with the expectation that miscellaneously found phenotypic features consolidate on a pathway level, under the assumption that they are functionally linked and collectively affect the disease phenotype.

High-throughput DNA sequencing can yield erroneous data [46]. MS based proteomics experiments generate enormous datasets that need to be carefully assessed [47].

Biological pathway databases play an essential role in annotating protein-coding genes resulting from high-throughput profiling approaches. There are approximately 547 pathway database resources available as listed in PathGuide (http://www.pathguide.org/). Albeit there are several well curated and reliable pathway database resources as also described by our group [48], significant efforts have been taken to expand biological pathway coverage beyond any single pathway data source. This is frequently carried out by integrating different sources in order to build high quality integrative pathway models without sacrificing data quality. However, biological data integration from heterogeneous sources has been challenging due to variability at the syntactic and semantic level. Syntactic variability is due to heterogeneity of molecular feature and pathway data formats, representation schemas and retrieval methods. Semantic

variability is due to incompatible pathway names, signalling event representations and molecular identifiers. For example, different pathway databases may choose to provide information on post-translation modifications, interacting proteins within a complex, or cellular location. Hence all these limitations have inhibited the growth of high quality integrative pathway models [49–51].

Another issue that arises when aiming to integrate data from different omics platforms is that conflicting results can be obtained. For example in some muscle invasive tumours presented in [52], transcriptomics analysis proved that the mRNA level of EGFR is up-regulated, whereas proteomic analysis did not show differential expression at the protein level. One explanation for such discrepancy may be translational regulation.

In KEGG, biological pathway categorization is currently available for several human key cellular processes [13]. Mapping MIBC-specific features (corrected on the level of the induced subgraph utilizing protein interaction information) to KEGG and performing enrichment analysis provided a total of 15 pathways (4 novel and 11 cited in published studies). 68 of 122 literature-mined protein candidates of relevance in muscle-invasive bladder cancer were identified as members of the identified pathways. This enabled to comprehensively rank pathways allowing the shortlisting of terms being individually discussed in the specific context of MIBC.

We focused on expanding our knowledge on muscle invasive urothelial neoplasm affected at the molecular level by comprehensively mapping available molecular datasets to pathways to build an interactome network utilizing public domain data sources. By differentiating the pathways based on previously described pathways and novel ones we obtained 11 modules that were known in context of bladder cancer muscle-invasion and 4 novel pathways. In respect to the previously known urothelial bladder muscle-invasive carcinoma pathways, our analysis retrieved pathways such as MAPK signalling pathway, ErbB signalling pathway, cell-cycle pathways and VEGF signalling pathway, hence confirming the systems-level approach for the particular phenotype [3, 12, 53, 54].

On the other hand, the interactome results also retrieved significant pathways comprising of signalling pathways, cytoskeleton remodelling pathways and neuromuscular junctions. Three molecular pathways were highly significant from the analysis, namely regulation of actin cytoskeleton, neurotrophin signalling pathway and endocytosis.

Neurotrophins are a class of closely related proteins that control the function, survival and development of neurons and have the potential to activate tropomyosin-related kinase (Trk) family of receptors and down regulate tumour necrosis factor superfamily (p75NTR) through which PI3K/Akt, Ras/Raf/MAP kinase, NF-kappa B and Jun kinase signalling pathways are triggered. Trk-receptors with neurotrophin ligands have been identified as initiating tumour progression, and the signalling pathway neurotrophins-Trk has been reported as a target for therapeutic intervention in hormone-refractory prostate cancer (HRPC) and in human astrocytomas, and potentially could play a role in urothelial carcinoma [55–58]. Endocytic pathways represent multiple aberrations in human neoplasms by being tightly and bi-directionally connected to signalling pathways that could indicate malignant transformations of the tumours. One of its regulators, DAB2, has also been reported to be prominent in advanced stages of urothelial cancers, where a decreased expression of the molecule could be observed in metastatic stages, and has been associated with high probabilities of recurrence and bladder carcinoma mortality [59–61]. Deregulation of actin bound proteins, namely p38β, ATF3 and Rho family of small GTPases which are involved in cytoskeletal remodelling, causes aberrant cell motility that leads to the muscle-invasive and metastatic phenotypes in cancer [62–65]. Our analysis highlights the role of the cytoskeletal remodelling pathway that contains integrins, cadherins and adhesion proteins. The respective molecular pathways discussed above open new avenues

for further investigation of urothelial muscle-invasive carcinoma. One enriched pathway that did not show any direct relation to bladder cancer was serotonergic synapse that contained 12 protein molecules (S3 Table).

The bioinformatics approach reported here involved integrating available public domain data sets in context of bladder muscle-invasive carcinoma on an interaction network, and further mapping them to biological pathway sources to reveal 15 pathways as being affected in progressive disease. Eleven from these pathways were discussed previously in the context of MIBC. It should be taken into account that while using such computational techniques to integrate molecular signatures from varying resources, certain technical issues regarding the use of appropriate global identifier need to be considered. In our approach, we discarded metabolite and micro-RNA targets for the pathway enrichment analysis (i.e. gene symbols mapped from metabolomics and miRNA data, service provided by HMDB and miRBase), resulting in 592 features from the total of 1,054 protein coding genes. This is mainly driven by hampered translation of metabolite and microRNA profiles to the level of involved protein coding genes, be it on the target or enzyme level. In regard to genomics and epigenetics, we only incorporated those gene symbols that contained epigenetic information on the protein/mRNA abundance levels for the interactome analysis.

On the other hand, the two pathway terms GnHR receptor and T cell receptor signalling pathways found as enriched on the basis of the 72 gene symbols being multiply identified in, were not retrieved from the analysis resting on the full set of 286 features being derived from the induced subgraph. Data evidence and selection biases clearly affect results of such integrated analysis demanding strict quality control of input data sets as followed in our study.

Apparently, each individual functional context highlights specific aspects of bladder cancer pathophysiology, but only providing limited characterization of clinical outcome on the cohort level.

In summary, automated data retrieval from the literature resulted in a first collection of molecular features associated with MIBC, and, complementing with omics profiling data, allowed augmenting a mechanistic (pathway) map linked to MIBC. From the cross-sectional nature of the underlying molecular feature collection no direct conclusion can be drawn regarding the prognostic relevance of individual pathways. However, deriving bladder cancer-associated protein coding genes on the basis of such pathway maps provides a systematic foundation for experimental analysis regarding association with development of muscle-invasive disease. We are confident that this approach can form the basis to rational selection of biomarkers for enabling targeted analysis of potentially relevant key molecules.

Conclusions

Our results suggest that there is a complex interplay between interacting pathways that characterizes the muscle invasive phenotype of invasive bladder cancer. We developed an integrated molecular model of muscle invasive bladder cancer to allow selecting protein-coding genes on the pathway level aimed at capturing a set of pathways of potential relevance in tumour progression. Further experimental validation of Neurotrophin signalling pathway, Regulation of actin cytoskeleton and Endocytosis with respect to disease progression and treatment response in muscle-invasive bladder carcinoma is indicated.

Supporting Information

S1 Table. Overlap Analysis, molecular feature sets from literature mining and omics screening. Sheet one lists protein coding genes retrieved from DNA mutation, methylation,

transcriptomics, metabolomics and literature screening. Redundant entries were ranked based on the frequency of occurence. Combing all protein coding genes resulted in 1,054 unique protein coding genes.
(XLS)

S2 Table. Data Inclusion for Muscle Invasive Bladder Carcinoma—Non-Redundant Features from Literature Mining. Provided are gene symbols pertaining to the muscle-invasive phenotype (T2–T4) of bladder carcinoma identified in scientific literature, together with expression levels and PubMed identifiers.
(XLS)

S3 Table. Enriched Pathways in context of Urothelial Muscle-Invasive Carcinoma using the 286 molecular features from the induced subgraph. Provided are KEGG pathways together with the number of MIBC-associated features, overlapping literature-mined gene symbols and Bonferonni corrected p-values. The supporting table is divided into two sheets namely Novel pathways and Literature-known pathways.
(XLS)

S4 Table. Data Inclusion for Muscle-Invasive Bladder Carcinoma—DNA-Methylation. Provided are gene symbols pertaining to the muscle-invasive phenotype (T2–T4) of bladder carcinoma identified from methylation studies, together with expression and methylation levels in addition to their PubMed identifiers.
(XLSX)

S5 Table. Data Inclusion for Muscle-Invasive Bladder Carcinoma—mRNA. Provided are gene symbols pertaining to the muscle-invasive phenotype (T2–T4) of bladder carcinoma resulting from transcriptomics studies (inculsion criteria of >50 molecules in the study), together with expression levels and PubMed identifiers. Protein coding genes that only hold differential expression information are provided with links from available studies.
(XLS)

S6 Table. Data Inclusion for Muscle-Invasive Bladder Carcinoma—miRNA. Provided are miRNA targets pertaining to the muscle-invasive phenotype (T2–T4) of bladder carcinoma from transcriptomics studies (inculsion criteria of >50 molecules in the study), together with expression levels and PubMed identifiers.
(XLS)

S7 Table. Data Inclusion for Muscle-Invasive Bladder Carcinoma—Metabolites. Provided are underlying enzymes for metabolites pertaining to the muscle-invasive phenotype (T2–T4) of bladder carcinoma from metabolomics studies (inculsion criteria of >50 molecules in the study), together with expression levels and PubMed identifiers. Protein coding genes that only hold differential expression information are provided with links to the data retrieved from available studies.
(XLS)

S8 Table. Data Inclusion for Muscle-Invasive Bladder Carcinoma—DNA Mutation. Provided are protein coding genes with significant levels of mutations pertaining to muscle-invasive phenotype (T2–T4) of bladder carcinoma and PubMed identifiers.
(XLS)

S9 Table. Protein coding genes involved in the induced Subgraph. Provided are the MIBC-asociated gene symbols in three different columns; a. 1,054 unique gene symbols from initial consolidation, b. 592 gene symbols excluding enzymes from metabolite signatures as well as

miRNA targets c. 286 gene symbols that formed the induced subgraph, and d. interaction information among the 286 gene symbols.
(XLS)

S10 Table. Enriched Pathways in context of Urothelial Muscle-Invasive Carcinoma using the 72 molecules that were found in more than one omics study. 7 pathway terms were enriched. The set of protein coding genes included is listed in sheet 2. Sheet 3 and 4 list known and novel pathways together with the number of protein coding genes assigned and Bonferroni corrected p-values. Sheet 5 and 6 list the overlapping gene symbols for all identified pathways.
(XLS)

Author Contributions

Conceived and designed the experiments: HM BM HH AB VJ. Performed the experiments: AB. Analyzed the data: AB. Contributed reagents/materials/analysis tools: AH PP IM JZ. Wrote the paper: AB AV JS AM. Designed and coordinated the study: HM BM. Provided support in data retrieval and subgraph computation: AH IM PP. Performed the analysis and drafted the manuscript: AB. Contributed to the interpretation of the results and drafted the publication along with reading and approving the final manuscript: AB AH BM PP IM HH AM JZ AV JS HM VJ.

References

1. Siegel R, Naishadham D, Jemal A (2013) Cancer statistics, 2013. CA Cancer J Clin 63: 11–30. doi: 10.3322/caac.21166 PMID: 23335087
2. Vishnu P, Mathew J, Tan WW (2011) Current therapeutic strategies for invasive and metastatic bladder cancer. Onco Targets Ther 4: 97–113. doi: 10.2147/OTT.S22875 PMID: 21792316
3. Mitra AP, Cote RJ (2009) Molecular pathogenesis and diagnostics of bladder cancer. Annu Rev Pathol 4: 251–285. doi: 10.1146/annurev.pathol.4.110807.092230 PMID: 18840072
4. Henning A, Wehrberger M, Madersbacher S, Pycha A, Martini T, et al. (2013) Do differences in clinical symptoms and referral patterns contribute to the gender gap in bladder cancer? BJU Int 112: 68–73. doi: 10.1111/j.1464-410X.2012.11661.x PMID: 23320798
5. Raman JD, Shariat SF, Karakiewicz PI, Lotan Y, Sagalowsky AI, et al. (2011) Does preoperative symptom classification impact prognosis in patients with clinically localized upper-tract urothelial carcinoma managed by radical nephroureterectomy? Urol Oncol 29: 716–723. doi: 10.1016/j.urolonc.2009.11.007 PMID: 20056458
6. Di Pierro GB, Gulia C, Cristini C, Fraietta G, Marini L, et al. (2012) Bladder cancer: a simple model becomes complex. Curr Genomics 13: 395–415. doi: 10.2174/138920212801619232 PMID: 23372425
7. Xiao J, Yiqing Q (2013) Bioinformatics analysis of the gene expression profile in Bladder carcinoma. Genet Mol Biol 36: 287–291. doi: 10.1590/S1415-47572013005000015 PMID: 23885213
8. Cauberg Evelyne CC, de la Rosette JJMCH, de Reijke TM (2011) Emerging optical techniques in advanced cystoscopy for bladder cancer diagnosis: A review of the current literature. Indian J Urol 27: 245–251. doi: 10.4103/0970-1591.82845 PMID: 21814317
9. Emmert-Streib F, Abogunrin F, de Matos Simoes R, Duggan B, Ruddock MW, et al. (2013) Collectives of diagnostic biomarkers identify high-risk subpopulations of hematuria patients: exploiting heterogeneity in large-scale biomarker data. BMC Med 11: 12. doi: 10.1186/1741-7015-11-12 PMID: 23327460
10. Ru Y, Dancik GM, Theodorescu D (2011) Biomarkers for prognosis and treatment selection in advanced bladder cancer patients. Curr Opin Urol 21: 420–427. doi: 10.1097/MOU.0b013e32834956d6 PMID: 21814055
11. Mitra AP, Datar RH, Cote RJ (2005) Molecular staging of bladder cancer. BJU Int 96: 7–12. doi: 10.1111/j.1464-410X.2005.05557.x PMID: 15963111
12. Ewald JA, Downs TM, Cetnar JP, Ricke WA (2013) Expression microarray meta-analysis identifies genes associated with Ras/MAPK and related pathways in progression of muscle-invasive bladder transition cell carcinoma. PLoS One 8: e55414. doi: 10.1371/journal.pone.0055414 PMID: 23383328
13. Mayer P, Mayer B, Mayer G (2012) Systems biology: building a useful model from multiple markers and profiles. Nephrol Dial Transplant 27: 3995–4002. doi: 10.1093/ndt/gfs489 PMID: 23144070

14. Kreeger PK, Lauffenburger DA (2010) Cancer systems biology: a network modeling perspective. Carcinogenesis 31: 2–8. doi: 10.1093/carcin/bgp261 PMID: 19861649
15. Besaratinia A, Cockburn M, Tommasi S (2013) Alterations of DNA methylome in human bladder cancer. Epigenetics 8. doi: 10.4161/epi.25927 PMID: 23975266
16. Chihara Y, Kanai Y, Fujimoto H, Sugano K, Kawashima K, et al. (2013) Diagnostic markers of urothelial cancer based on DNA methylation analysis. BMC Cancer 13: 275. doi: 10.1186/1471-2407-13-275 PMID: 23735005
17. Weiss GJ, Liang WS, Demeure MJ, Kiefer JA, Hostetter G, et al. (2013) A pilot study using next-generation sequencing in advanced cancers: feasibility and challenges. PLoS One 8: e76438. doi: 10.1371/journal.pone.0076438 PMID: 24204627
18. Mischak H, Schanstra JP (2011) CE-MS in biomarker discovery, validation, and clinical application. Proteomics Clin Appl 5: 9–23. doi: 10.1002/prca.201000058 PMID: 21280234
19. Alberice JV, Amaral AFS, Armitage EG, Lorente JA, Algaba F, et al. (2013) Searching for urine biomarkers of bladder cancer recurrence using a liquid chromatography-mass spectrometry and capillary electrophoresis-mass spectrometry metabolomics approach. J Chromatogr A 1318: 163–170. doi: 10.1016/j.chroma.2013.10.002 PMID: 24139504
20. GeneCards (2013) The Human Gene Compendium. Crown Hum Genome Center, Dep Mol Genet Weizmann Inst Sci.
21. Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, et al. (2004) ONCOMINE: a cancer microarray database and integrated data-mining platform. Neoplasia 6: 1–6. doi: 10.1016/S1476-5586(04)80047-2 PMID: 15068665
22. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res 33: D514–D517. doi: 10.1093/nar/gki033 PMID: 15608251
23. Mayer G, Heinze G, Mischak H, Hellemons ME, Heerspink HJL, et al. (2011) Omics-bioinformatics in the context of clinical data. Methods Mol Biol 719: 479–497. doi: 10.1007/978-1-61779-027-0_22 PMID: 21370098
24. Abu-Asab MS, Chaouchi M, Alesci S, Galli S, Laassri M, et al. (2011) Biomarkers in the age of omics: time for a systems biology approach. OMICS 15: 105–112. doi: 10.1089/omi.2010.0023 PMID: 21319991
25. De Las Rivas J, Fontanillo C (2010) Protein-protein interactions essentials: key concepts to building and analyzing interactome networks. PLoS Comput Biol 6: e1000807. doi: 10.1371/journal.pcbi.1000807 PMID: 20589078
26. Wixon J, Kell D (2000) The Kyoto encyclopedia of genes and genomes–KEGG. Yeast 17: 48–55.
27. Mi H, Muruganujan A, Thomas PD (2013) PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. Nucleic Acids Res 41: D377–86. doi: 10.1093/nar/gks1118 PMID: 23193289
28. Croft D, O'Kelly G, Wu G, Haw R, Gillespie M, et al. (2011) Reactome: a database of reactions, pathways and biological processes. Nucleic Acids Res 39: D691–D697. doi: 10.1093/nar/gkq1018 PMID: 21067998
29. Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, et al. (2009) AmiGO: online access to ontology and annotation data. Bioinformatics 25: 288–289. doi: 10.1093/bioinformatics/btn615 PMID: 19033274
30. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, et al. (2013) NCBI GEO: archive for functional genomics data sets–update. Nucleic Acids Res 41: D991–5. doi: 10.1093/nar/gks1193 PMID: 23193258
31. Rustici G, Kolesnikov N, Brandizi M, Burdett T, Dylag M, et al. (2013) ArrayExpress update–trends in database growth and links to data analysis tools. Nucleic Acids Res 41: D987–90. doi: 10.1093/nar/gks1174 PMID: 23193272
32. Maglott D, Ostell J, Pruitt KD, Tatusova T (2007) Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res 35: D26–31. doi: 10.1093/nar/gkl993 PMID: 17148475
33. Kerrien S, Aranda B, Breuza L, Bridge A, Broackes-Carter F, et al. (2012) The IntAct molecular interaction database in 2012. Nucleic Acids Res 40: D841–6. doi: 10.1093/nar/gkr1088 PMID: 22121220
34. Stark C, Breitkreutz B-J, Chatr-Aryamontri A, Boucher L, Oughtred R, et al. (2011) The BioGRID Interaction Database: 2011 update. Nucleic Acids Res 39: D698–D704. doi: 10.1093/nar/gkq1116 PMID: 21071413
35. Bindea G, Mlecnik B, Hackl H, Charoentong P, Tosolini M, et al. (2009) ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics 25: 1091–1093. doi: 10.1093/bioinformatics/btp101 PMID: 19237447

36. Bindea G, Galon J, Mlecnik B (2013) CluePedia Cytoscape plugin: pathway insights using integrated experimental and in silico data. Bioinformatics 29: 661–663. doi: 10.1093/bioinformatics/btt019 PMID: 23325622

37. Huang H, McGarvey PB, Suzek BE, Mazumder R, Zhang J, et al. (2011) A comprehensive protein-centric ID mapping service for molecular data integration. Bioinformatics 27: 1190–1191. doi: 10.1093/bioinformatics/btr101 PMID: 21478197

38. Wishart DS, Tzur D, Knox C, Eisner R, Guo AC, et al. (2007) HMDB: the Human Metabolome Database. Nucleic Acids Res 35: D521–6. doi: 10.1093/nar/gkl923 PMID: 17202168

39. Kozomara A, Griffiths-Jones S (2011) miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res 39: D152–7. doi: 10.1093/nar/gkq1027 PMID: 21037258

40. Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ (2008) miRBase: tools for microRNA genomics. Nucleic Acids Res 36: D154–8. doi: 10.1093/nar/gkm952 PMID: 17991681

41. Bernthaler A, Mühlberger I, Fechete R, Perco P, Lukas A, et al. (2009) A dependency graph approach for the analysis of differential gene expression profiles. Mol Biosyst 5: 1720–1731. doi: 10.1039/b903109j PMID: 19585005

42. Rapberger R, Perco P, Sax C, Pangerl T, Siehs C, et al. (2008) Linking the ovarian cancer transcriptome and immunome. BMC Syst Biol 2: 2. doi: 10.1186/1752-0509-2-2 PMID: 18173842

43. Mischak H, Allmaier G, Apweiler R, Attwood T, Baumann M, et al. (2010) Recommendations for biomarker identification and qualification in clinical proteomics. Sci Transl Med 2: 46ps42. doi: 10.1126/scitranslmed.3001249 PMID: 20739680

44. Vlahou A (2013) Network views for personalized medicine. Proteomics Clin Appl 7: 384–387. doi: 10.1002/prca.201200121 PMID: 23532915

45. Frantzi M, Zerefos PG, Sanchez-Carbayo M, Anagnou NP, Zoidakis J, et al. (2012) Profilin 1 is a Potential Biomarker for Bladder Cancer Aggressiveness. Mol Cell Proteomics 11: M111.009449–M111.009449. doi: 10.1074/mcp.M111.009449 PMID: 22159600

46. Kircher M, Kelso J (2010) High-throughput DNA sequencing–concepts and limitations. Bioessays 32: 524–536. doi: 10.1002/bies.200900181 PMID: 20486139

47. Dakna M, He Z, Yu WC, Mischak H, Kolch W (2009) Technical, bioinformatical and statistical aspects of liquid chromatography-mass spectrometry (LC-MS) and capillary electrophoresis-mass spectrometry (CE-MS) based clinical proteomics: a critical assessment. J Chromatogr B Analyt Technol Biomed Life Sci 877: 1250–1258. doi: 10.1016/j.jchromb.2008.10.048 PMID: 19010091

48. Frantzi M, Bhat A, Latosinska A (2014) Clinical proteomic biomarkers: relevant issues on study design & technical considerations in biomarker development. Clin Transl Med 3: 7. doi: 10.1186/2001-1326-3-7 PMID: 24679154

49. Krishnamurthy L, Nadeau J, Ozsoyoglu G, Ozsoyoglu M, Schaeffer G, et al. (2003) Pathways database system: an integrated system for biological pathways. Bioinformatics 19: 930–937. doi: 10.1093/bioinformatics/btg113 PMID: 12761054

50. Stein LD (2003) Integrating biological databases. Nat Rev Genet 4: 337–345. doi: 10.1038/nrg1065 PMID: 12728276

51. Philippi S, Köhler J (2006) Addressing the problems with life-science databases for traditional uses and systems biology. Nat Rev Genet 7: 482–488. doi: 10.1038/nrg1872 PMID: 16682980

52. Comprehensive molecular characterization of urothelial bladder carcinoma. (2014). Nature 507: 315–322. doi: 10.1038/nature12965 PMID: 24476821

53. Wu X-R (2005) Urothelial tumorigenesis: a tale of divergent pathways. Nat Rev Cancer 5: 713–725. doi: 10.1038/nrc1697 PMID: 16110317

54. McConkey DJ, Lee S, Choi W, Tran M, Majewski T, et al. (2010) Molecular genetics of bladder cancer: Emerging mechanisms of tumor initiation and progression. Urol Oncol 28: 429–440. doi: 10.1016/j.urolonc.2010.04.008 PMID: 20610280

55. Reichardt LF (2006) Neurotrophin-regulated signalling pathways. Philos Trans R Soc Lond B Biol Sci 361: 1545–1564. doi: 10.1098/rstb.2006.1894 PMID: 16939974

56. Thiele CJ, Li Z, McKee AE (2009) On Trk–the TrkB signal transduction pathway is an increasingly important target in cancer biology. Clin Cancer Res 15: 5962–5967. doi: 10.1158/1078-0432.CCR-08-0651 PMID: 19755385

57. Koochekpour S (2010) Androgen receptor signaling and mutations in prostate cancer. Asian J Androl 12: 639–657. doi: 10.1038/aja.2010.89 PMID: 20711217

58. Assimakopoulou M, Kondyli M, Gatzounis G, Maraziotis T, Varakis J (2007) Neurotrophin receptors expression and JNK pathway activation in human astrocytomas. BMC Cancer 7: 202. doi: 10.1186/1471-2407-7-202 PMID: 17971243

59. Mosesson Y, Mills GB, Yarden Y (2008) Derailed endocytosis: an emerging feature of cancer. Nat Rev Cancer 8: 835–850. doi: 10.1038/nrc2521 PMID: 18948996

60. Polo S, Pece S, Di Fiore PP (2004) Endocytosis and cancer. Curr Opin Cell Biol 16: 156–161. doi: 10.1016/j.ceb.2004.02.003 PMID: 15196558

61. Karam JA, Shariat SF, Huang H-Y, Pong R-C, Ashfaq R, et al. (2007) Decreased DOC-2/DAB2 expression in urothelial carcinoma of the bladder. Clin Cancer Res 13: 4400–4406. doi: 10.1158/1078-0432.CCR-07-0287 PMID: 17671122

62. Price LS, Collard JG (2001) Regulation of the cytoskeleton by Rho-family GTPases: implications for tumour cell invasion. Semin Cancer Biol 11: 167–173. doi: 10.1006/scbi.2000.0367 PMID: 11322835

63. Yu L, Yuan X, Wang D, Barakat B, Williams ED, et al. (2014) Selective regulation of p38β protein and signaling by integrin-linked kinase mediates bladder cancer cell migration. Oncogene 33: 690–701. doi: 10.1038/onc.2013.20 PMID: 23435415

64. Yuan X, Yu L, Li J, Xie G, Rong T, et al. (2013) ATF3 suppresses metastasis of bladder cancer by regulating gelsolin-mediated remodeling of the actin cytoskeleton. Cancer Res 73: 3625–3637. doi: 10.1158/0008-5472.CAN-12-3879 PMID: 23536558

65. Kamai T, Tsujii T, Arai K, Takagi K, Asami H, et al. (2003) Significant association of Rho/ROCK pathway with invasion and metastasis of bladder cancer. Clin Cancer Res 9: 2632–2641. PMID: 12855641

---------- Forwarded message ----------
From: <bioinfopro@jacobspublishers.international>
Date: 4 November 2015 at 11:07
Subject: Article accepted for publication
To: Akshay Bhat <akshayb04@gmail.com>
Cc: izoidakis@bioacademy.gr

Dear Dr. Jerome Zoidakis / Akshay Bhat

Warm greetings from Jacobs Publishers!

Hope everything is going fine.

Congratulations your article has been accepted for publication.

Soon you will receive to final proof of the article.

We would also like to inform you that you will receive an invoice of 499 USD from accounts department
after you receive the final proof of the article.

Waiting for your valuable response regarding this.

Thank you have a great day.

Regards

Issac Jonan
Jacobs Journal of Bioinformatics and Proteomics
Jacobs Publishers
9600 Great Hills
Trail # 150 w
Austin, Texas
78759(Travis County)
E-mail: bioinfopro@jacobspublishers.international

Comparison of ClueGO and ImPAla for integrated pathway enrichment analysis

Akshay Bhat[1, 2], Vera Jankowski[3], Antonia Vlahou[4], Harald Mischak[2, 5], Jerome Zoidakis[4]

[1]Charité-Universitätsmedizin Berlin, Med. Klinik IV, Berlin, Germany

[2]Mosaiques diagnostics GmbH, Hannover, Germany

[3]Institute for Molecular Cardiovascular Research (IMCAR), Aachen, Germany

[4]Biomedical Research Foundation Academy of Athens, Biotechnology Division, Athens, Greece

[5]BHF Glasgow Cardiovascular Research Centre, University of Glasgow, Glasgow, UK

Corresponding author:

Jerome Zoidakis PhD.

Biomedical Research Foundation, Academy of Athens

Department of Biotechnology

Soranou Efessiou 4

11527 Athens, Greece

Tel. 30-210-6597485

E-Mail: izoidakis@bioacademy.gr

Abstract

Background: High-throughput experimental technologies ranging from genomic sequencing and gene/protein profiling are now commonly being used for the molecular characterization of diseases. These techniques produce large datasets of differentially expressed features that include genes, mRNAs, proteins and metabolites. The abundant data defy straightforward intuitive interpretation. Hence the correlation of molecular features to biological pathways may ultimately help in understanding the patho-physiology of a disease. Many available computational tools allow annotating such integrated datasets at the pathway level. Two prominent tools are ClueGO (Cytoscape plug-in) and ImPAla (web based application). Both tools provide advantages in integrating different pathway databases. However, each tool abides by specific statistical and mathematical algorithms in enriching molecular features onto pathway-centric networks.

Materials and Methods: Bladder cancer (BC) specific molecular features were retrieved from literature and omics profiles. The data comprise of differentially expressed DNA-mutations, DNA-methylation, mRNAs, miRNAs, proteins and metabolites. These features were combined and subjected to protein-protein interactions to yield the BC interactome. The features from this interactome were used as the input-list for the pathway enrichment analysis.

Results: 292 pathways were obtained from ClueGO and 471 pathways from ImPAla. The resulting pathways were selected according to the following significance criterion: multiple comparison corrected p-value <0.05. Comparison of the results obtained by the two applications yielded 152 pathway terms with exactly the same name. Moreover, 137 ClueGO pathway terms were similar to 251 ImPAla pathways. Thus, the overall overlap between the two datasets is 289 ClueGO pathways corresponding to 403 ImPAla pathways. ClueGO yielded 3 unique pathway terms whereas in the case of ImPAla 68 unique pathways were obtained. Both datasets contain redundant terms but the ImPAla results are characterized by higher redundancy. In addition, ImPAla yields 12 unique pathways that are not related to BC.

Conclusion: Cytoscape-ClueGO has better performance than ImPAla in pathway enrichment analysis since the output is less redundant and contains all the biologically significant information.

Keyword: bioinformatics, systems biology, ClueGO, ImPAla, data integration, pathway enrichment

Introduction

Next-generation sequencing and profiling techniques ranging from genomics to transcriptomics, proteomics and metabolomics have transformed biological research by allowing a comprehensive monitoring of biological systems [1]. These technologies yield a vast amount of data, typically as a list of differentially expressed proteins, genes, transcripts, miRNA and metabolites that may have specific roles in a given clinical phenotype [2]. However, these lists of individual features fail in providing a mechanistic insight for the molecular characterization of a disease [3]. Hence, these challenges have led to an advent of new functional annotation approaches in which individual features are grouped together into pathways by statistical or mathematical algorithms [2]. An important advantage of working with molecular pathways rather than individual proteins or genes is the fact that it is often easier and more relevant to predict the function of a module than a function of an individual protein/gene [3]. Prediction of a functional module is possible only if the pathway contains a sufficient number of features known to be associated to that pathway. Such functional module prediction is also known as enrichment analysis; as it builds on the assumption that features could be assigned to a particular pathway or process, grouped and organized in Gene Ontologies (GO) [4]. GO are sub-categorized into cellular component, biological process and molecular pathway. Enrichment analysis determines whether the number of features attributed to a specific pathway is higher than expected by chance. This can be calculated using statistical methods such as $\chi 2$, hypergeometric tests and Fisher's exact tests and have been implemented in many software packages, like R-Project (http://www.r-project.org/). Some of the frequently used software packages and applications for performing enrichment analysis are publically available. Some of the packages include ClueGO [5], BinGO [6], Gorilla [7], Enrichment Map [8], Metscape [9], InCroMap [10], 3Omics [11], iPEAP [12] and ImPAla [13]. The aim of this short communication is to evaluate the performance of pathway enrichment analysis obtained from two bioinformatics tools ImPAla and ClueGO. The two applications were selected because ClueGO has more than 120 citations

(compared to other programs) whereas 18 articles cited ImPAla (a significantly higher number of citations compared to other web-based enrichment tools.

Materials and Methods

The list of bladder cancer associated features that were subjected to protein-protein interactions (PPIs) was obtained from the BcCluster database (www.bccluster.org) [14]. Protein interaction information was acquired by querying PPI databases such as IntAct [15], BioGRID [16], String [17] and Reactome [18]. First, all available PPIs for the human proteome were downloaded into Cytoscape [19] to form the human interactome. The PPIs relevant to the BC-associated proteins were retrieved from this human interactome. Only proteins that had at least one binding partner were retained. This step yielded 435 entries (in official gene-ids) that correspond to proteins from the BC PPI. The list of 435 entries with information regarding the official gene id and gene name, in addition to the differential expression of the feature is provided in Supplementary Table S1. This list was then subjected to pathway enrichment using ClueGO and ImPAla. ClueGO provides an advantage to perform cluster comparisons for pathway enrichment and allows the option to separately input up and down regulated molecules in the software. In addition, ClueGO provides an optional redundancy reduction feature (Fusion) to assess GO terms that share similar associated features in a parent–child relation. This option was selected in our ClueGO pathway enrichment analysis to eliminate the redundant pathway terms. In contrast, ImPAla does not provide an option of redundancy reduction for pathway terms. The pathway databases selected for enrichment were KEGG [20] and Reactome. The statistical selection criterion taken into account for the enrichment analysis was the corrected for multiple comparisons p-value < 0.05. The overlap assessment between the pathway outputs was performed manually.

Results and Discussion

The overview of the pathway analysis performed by ClueGO and ImPAla is illustrated in Table 1. The total number of KEGG and Reactome pathway terms obtained from ClueGO was 292. ImPAla produced 471 pathways. Additional information for the raw enriched pathway results is also provided in Supplementary Table S2. This information includes number of input genes, corrected p-value scores, Gene Ontology source and the number of genes held in each pathway. By comparing the pathway results, 152 pathway terms exactly overlapped in ClueGO and ImPAla. 137 pathway terms from ClueGO were highly similar to 251 ImPAla pathway terms. Therefore, the total calculated overlap of pathways between the two tools equalled to 289 ClueGO pathways that correspond to 403 ImPAla enriched pathways. In addition, the software programs also produced unique pathway terms. There were 3 unique pathways from the total 292 ClueGO pathway terms whereas 68 pathways were unique from ImPAla. Both the enrichment tools yielded redundancy in the output results, however results from ImPAla were characterized by higher redundancies in pathway terms (for e.g. the pathway terms “DNA replication”, “synthesis of DNA”). Moreover, from the unique set of 68 ImPAla pathway terms, 12 pathway terms were not related in the context of BC. Some of these pathways include alcoholism, amphetamine addiction, inflammatory bowel disease (IBD), malaria, viral myocarditis and prion diseases. On the contrary, the 3 unique pathways obtained by ClueGO were relevant to BC. It was also noted that the overlapping pathway terms from ImPAla and ClueGO contained pathway names that are not relevant in the context of BC. These common terms totalled to 34 ImPAla and 30 ClueGO pathway terms. The common pathway terms included oocyte meiosis, tuberculosis, type II diabetes mellitus, circadian clock and shigellosis. In addition, there were some common terms that were very general in the description, for e.g. “Disease and Developmental biology”. The resulting raw pathway outputs, exactly overlapping pathways, highly similar pathways, unique pathways and unrelated pathways retrieved from the two programs are provided in Supplementary Table S2. In addition, the comparison of significant

overlapping pathways obtained from ClueGO and ImPAla is represented as a Venn diagram in Figure 1.

Software	Availability	User input	p-value correction method	Total pathway output	Reference
ClueGO	Cytoscape plugin	435 entries	Bonferroni	292	[5]
ImPAla	Web-based	435 entries	Benjamini Hochberg	471	[13]

Table 1. General information for the results obtained from the pathway enrichment analysis.

In conclusion, output by both software tools provided a significant set of pathways for the enrichment analysis. However, ImPAla produced a significantly higher number of pathways than ClueGO. This is due to the fact that ImPAla does not have a pathway term redundancy reduction feature. Hence this could be the reason for the high redundancy observed in the ImPAla pathway term list (many common pathways with different names). For e.g. the pathway term Chagas disease (American trypanosomiasis) was retrieved from ClueGO, whereas Chagas disease (American trypanosomiasis) and African trypanosomiasis, were retrieved from ImPAla. In our previous publication [14], we had filtered 292 ClueGO pathways to make them non-redundant for the database storage. This non-redundant pathway list equals 90 BC specific pathways. ImPAla retrieves more redundant pathways than ClueGO. Thus, the effort to manually eliminate redundant terms from the 471 ImPAla derived pathways is significantly higher. In addition, the 68 unique pathways retrieved by ImPAla contain 12 pathways not related to BC. Both pathway enrichment tools allow the input of regulation information as numerical values (fold change) and p-values. However, ClueGO has the additional feature of allowing text input for regulation. Since our data set contained regulation information in the form of text (up/down) (Table S1), ClueGO was able to incorporate this feature in the pathway enrichment analysis. It should be noted that if the user in ImPAla does not provide numerical

regulation values, the tool considers the protein/gene as differentially expressed but does not assign a specific trend (up or down). ClueGO provides additional options that allow the user to define the stringency of pathway selection. The options include: Kappa statistics in order to generate pathway network visualizations, mid-p-values and doubling p-values in order to retrieve significant pathways based on user-defined threshold p-values, setting specific limits for ratios of differentially expressed genes relative to the total number of genes present in a pathway in order to consider a pathway as significant. In contrast, ImPAla does not provide these features. Moreover, ClueGO provides a more descriptive data output that contains significant additional information when compared to ImPAla data output. The common columns shared among the two enrichment tools include pathway name, pathway source, number of input genes present in the pathway, total number of genes present in the pathway and corrected p-values. Additional columns provided by ClueGO were genes down-regulated in input set, genes up-regulated in input set, Gene Ontology ID, and percentage input Genes present in pathway. In addition, ClueGO is more user-friendly when compared to ImPAla since it offers better help options.

In the study by Jaakkola MK et al. the performance of six enrichment tools were tested on experimental datasets from six renal-cell carcinoma and four type-1 diabetes samples. The software programs tested included, SPIA, CePa, DAVID, NetGSA, GSEA and Pathifier. From the resulting enrichment outputs, the authors noted that significant pathways were different according to different enrichment methods, and the number of significant findings depended on the enrichment method. Hence, they conclude that the selection of the enrichment method had a large impact on the pathway output results [21]. Our attempt was to compare the performance for significant pathway outputs yielded from ClueGO and ImPAla in the context to bladder cancer. It could be stated that ImPAla can provide advantages over ClueGO by integrating many more pathway database resources for the comprehensiveness in pathway information. However, we only selected two widely used and up-to-date pathway database resources, KEGG and Reactome. Adding more databases in our analysis would introduce higher redundancy in pathway outputs. In regard to manually updating database

sources, the ClueGO application allows users to update individual pathway database source within Cytoscape in order to obtain latest data whereas ImPAla is an omics-integration (with a focus towards metabolomics integration) and pathway enrichment application that contains the latest update of January 2015. In addition, ImPAla also allows the incorporation of differential expression information for molecules such as fold change and p-values. Nevertheless, ImPAla does not offer the option to input separately up and down regulated genes and does not make predictions on the activation/deactivation of an affected pathway in contrast to ClueGO. In addition, Cytoscape provides various plug-ins for analysing different omics datasets such as genes, mRNAs, proteins, SNPs, metabolites and miRNAs. This gives the advantage in using one analysis and visualization tool for all high-throughput sequencing and profiling experiments. Furthermore, having a single analysis tool also helps to prevent errors due to compatibility when transferring data between different software applications. Therefore, we conclude that Cytoscape-ClueGO is preferable to ImPAla for pathway enrichment and in the comprehensive characterization of molecular diseases.

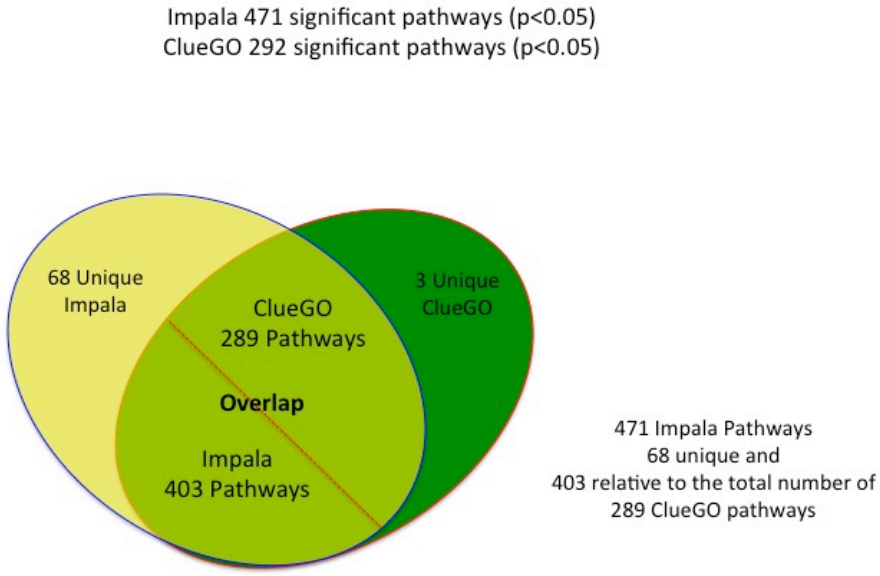

Figure 1. Venn Diagram represents the overlap of pathway terms between ClueGO and ImPAla. All pathways are selected based on p-value < 0.05

Acknowledgement

HM, VJ, AV and JZ designed and coordinated the study, and AB performed data retrieval, the analysis and drafted the manuscript. All authors contributed to the interpretation of the results and drafted the publication along with reading and approving the final manuscript. This work was supported in part by the BCMolMed grant PITN-GA-2012-317450. BCMolMed is funded by the European Commission.

Reference

1. Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Comput. Biol. 2012;8:e1002375.

2. Bhat A, Heinzel A, Mayer B, Perco P, Mühlberger I, Husi H, et al. Protein interactome of muscle invasive bladder cancer. PLoS One. 2015;10:e0116404.

3. Bhat A, Dakna M, Mischak H. Integrating proteomics profiling data sets: a network perspective. Methods Mol. Biol. 2015;1243:237–53.

4. Blake JA, Dolan M, Drabkin H, Hill DP, Li N, Sitnikov D, et al. Gene Ontology annotations and resources. Nucleic Acids Res. 2013;41:D530–5.

5. Bindea G, Mlecnik B, Hackl H, Charoentong P, Tosolini M, Kirilovsky A, et al. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics. 2009;25:1091–3.

6. Maere S, Heymans K, Kuiper M. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics. 2005;21:3448–9.

7. Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics. 2009;10:48.

8. Isserlin R, Merico D, Voisin V, Bader GD. Enrichment Map - a Cytoscape app to visualize and explore OMICs pathway enrichment results. F1000Research.2014;3:141.

9. Karnovsky A, Weymouth T, Hull T, Tarcea VG, Scardoni G, Laudanna C, et al. Metscape 2 bioinformatics tool for the analysis and visualization of metabolomics and gene expression data. Bioinformatics.2012;28:373–80.

10. Eichner J, Rosenbaum L, Wrzodek C, Häring H-U, Zell A, Lehmann R. Integrated enrichment analysis and pathway-centered visualization of metabolomics, proteomics, transcriptomics, and genomics data by using the InCroMAP software. J. Chromatogr. B. Analyt. Technol. Biomed. Life Sci. 2014;966:77–82.

11. Kuo T-C, Tian T-F, Tseng YJ. 3Omics: a web-based systems biology tool for analysis, integration and visualization of human transcriptomic, proteomic and metabolomic data. BMC Syst. Biol. 2013;7:64.

12. Sun H, Wang H, Zhu R, Tang K, Gong Q, Cui J, et al. iPEAP: integrating multiple omics and genetic data for pathway enrichment analysis. Bioinformatics. 2014;30:737–9.

13. Kamburov A, Cavill R, Ebbels TMD, Herwig R, Keun HC. Integrated pathway-

level analysis of transcriptomics and metabolomics data with IMPaLA. Bioinformatics. 2011;27:2917–8.

14. Bhat A, Mokou M, Zoidakis J, Jankowski V, Vlahou A, Mischak H. BcCluster: A bladder cancer database at the molecular level. Bl. Cancer. 2015.

15. Kerrien S, Aranda B, Breuza L, Bridge A, Broackes-Carter F, Chen C, et al. The IntAct molecular interaction database in 2012. Nucleic Acids Res. 2012;40:D841–6.

16. Stark C, Breitkreutz B-J, Chatr-Aryamontri A, Boucher L, Oughtred R, Livstone MS, et al. The BioGRID Interaction Database: 2011 update. Nucleic Acids Res. 2011;39:D698–704.

17. Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, et al. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res. 2013;41:D808–15.

18. Croft D, O'Kelly G, Wu G, Haw R, Gillespie M, Matthews L, et al. Reactome: a database of reactions, pathways and biological processes. Nucleic Acids Res. 2011;39:D691–7.

19. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13:2498–504.

20. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for deciphering the genome. Nucleic Acids Res. 2004;32:D277–80.

21. Jaakkola MK, Elo LL. Empirical comparison of structure-based pathway methods. Brief. Bioinform. 2015;bbv049.

Identification of ageing-associated naturally occurring peptides in human urine

Esther Nkuipou-Kenfack[1,2], Akshay Bhat[1,3], Julie Klein[4,5], Vera Jankowski[6], William Mullen[7], Antonia Vlahou[8,9], Mohammed Dakna[1], Thomas Koeck[1], Joost P. Schanstra[5,6], Petra Zürbig[1], Karl L. Rudolph[10], Björn Schumacher[11], Andreas Pich[2] and Harald Mischak[1,7]

[1] Mosaiques Diagnostics GmbH, Hannover, Germany

[2] Hannover Medical School, Core Facility Proteomics, Hannover, Germany

[3] Charité-Universitätsmedizin Berlin, Med. Klinik IV, Berlin, Germany

[4] Institut National de la Santé et de la Recherche Médicale (INSERM), Institut of Cardiovascular and Metabolic Disease, Toulouse, France

[5] Université Toulouse III Paul-Sabatier, Toulouse, France

[6] University Hospital RWTH Aachen, Institute for molecular cardiovascular research (IMCAR), Aachen, Germany

[7] BHF Glasgow Cardiovascular Research Centre, University of Glasgow, Glasgow, United Kingdom

[8] Biotechnology Division, Biomedical Research Foundation Academy of Athens, Athens, Greece

[9] School of Biomedical and Healthcare Sciences, Plymouth University, Plymouth, UK

[10] Leibniz Institute of Age Research, Fritz Lipmann Institute, Jena, Germany

[11] Institute for Genome Stability in Ageing and Disease and Cologne Excellence Cluster for Cellular Stress Responses in Aging-Associated Diseases (CECAD) Research Center, University of Cologne, Cologne, Germany

Correspondence to: *Petra Zürbig,* ***email***: *zuerbig@mosaiques-diagnostics.com*

Keywords: ageing, urine, peptidomics, collagen, proteases, systems biology, Gerotarget

Received: *September 10, 2015* ***Accepted***: *September 16, 2015* ***Published***: *September 29, 2015*

ABSTRACT

To assess normal and pathological peptidomic changes that may lead to an improved understanding of molecular mechanisms underlying ageing, urinary peptidomes of 1227 healthy and 10333 diseased individuals between 20 and 86 years of age were investigated. The diseases thereby comprised diabetes mellitus, renal and cardiovascular diseases. Using age as a continuous variable, 116 peptides were identified that significantly ($p < 0.05$; $|\rho| \geq 0.2$) correlated with age in the healthy cohort. The same approach was applied to the diseased cohort. Upon comparison of the peptide patterns of the two cohorts 112 common age-correlated peptides were identified. These 112 peptides predominantly originated from collagen, uromodulin and fibrinogen. While most fibrillar and basement membrane collagen fragments showed a decreased age-related excretion, uromodulin, beta-2-microglobulin and fibrinogen fragments showed an increase. Peptide-based *in silico* protease analysis was performed and 32 proteases, including matrix metalloproteinases and cathepsins, were predicted to be involved in ageing. Identified peptides, predicted proteases and patient information were combined in a systems biology pathway analysis to identify molecular pathways associated with normal and/or pathological ageing. While perturbations in collagen homeostasis, trafficking of toll-like receptors and endosomal pathways were commonly identified, degradation of insulin-like growth factor-binding proteins was uniquely identified in pathological ageing.

INTRODUCTION

Normal physiological ageing is a complex, multi-mechanistic systemic process that is influenced by genetic and environmental factors. It leads to a gradual decline in biological functions. Key molecular mechanisms identified in ageing include genomic instability, telomere attrition, loss of proteostasis and mitochondrial dysfunction [1]. However, information on normal physiological ageing may be blurred by alterations associated with pathologies (acute and chronic) developing in parallel with ageing and it is still often unclear whether an observed molecular change is due to ageing, or is (partially) due to concomitant diseases. It is thus obvious that more efforts should be invested into understanding molecular pathways underlying ageing in both healthy and diseased individuals. These may lead to strategies for the management of pathological complications during ageing.

As ageing is a complex systemic process, "omics" approaches aiming at studying multiple features at once, have been applied with the aim to unravel novel underlying molecular processes. Proteomic studies confirmed that oxidative stress occurs ubiquitously during ageing while other events were shown to be more tissue-specific (reviewed in [2]). However, a shortcoming in most of these studies was the use of animal models [2]. The scarcity of human studies can be largely attributed to the inability in obtaining appropriate tissue samples. Thus, a way forward could be the investigation of readily available body fluids.

In a first small scale study, we investigated the urinary proteome in a cohort of 324 healthy individuals between 2 to 73 years of age showing the feasibility to obtain high resolution molecular information from readily available body fluids such as urine [3]. Meanwhile, we have accumulated multiple high-resolution urine peptidomics datasets that enable the investigation of ageing-associated changes in a large cohort [4]. In the present study, we therefore investigated the unique urinary proteome profiles of 11560 individuals in an attempt to identify specific ageing-associated alterations and investigate pathological derailment of normal ageing. This showed that perturbations in collagen homeostasis, trafficking of toll-like receptors and endosomal pathways were associated to healthy ageing, while degradation of insulin-like growth factor-binding proteins was uniquely identified in pathological ageing

RESULTS

Age-correlation analysis in the healthy group

Among the 11560 individual urinary peptidomes, 1227 originated from individuals without disease and were thus considered healthy (age 20-86). Correlation analysis of 2223 individual sequenced peptides with age performed in the healthy peptidomes identified 116 significantly ageing-associated peptides ($p \leq 0.05$) (Supplemental Table 1). These peptides predominantly included fragments of collagen, fibrinogen, and uromodulin. Collagen fragments comprised 83 (72%) out of the 116 peptides identified. Amongst collagen fragments, most peptides originated from fibrillar collagens (89%) including type I collagen (47%) and type III collagen (11%) while basement membrane type IV collagens alpha-1and -3 showed a low abundance (2%).

The majority of peptides (65%) showed a negative correlation with age (Supplemental Table 1). The two most negatively age-correlated peptides were two type I collagen alpha-1 fragments ($\rho = -0.324$, $p < 0.0001$ and $\rho = -0.315$, $p < 0.0001$, Supplemental Table 1) and 93% of the type I collagen fragments decreased during ageing. Other negatively age-correlated peptides originated from 5-AMP-activated protein kinase subunit gamma-3 (PRKAG3), AMP/ATP-binding subunit of AMP-activated protein kinase (AMPK) and blood-derived proteins (beta-2-microglobulin, fibrinogen alpha and beta chains). Contrarily, the two most positively age-correlated peptides were type IV collagen alpha-3 and type II collagen alpha-1 fragments ($\rho = 0.504$, $p < 0.0001$ and $\rho = 0.451$, $p < 0.0001$ respectively, Supplemental Table 1). Additionally an age-dependent increase in almost 50% of type III collagen and 83% in type II collagen fragments was observed. Other positively age-correlated peptides originated from clusterin, haptoglogin, cystatin-B, retinol-binding protein 4, CD99 antigen, and the kidney-specific peptide uromodulin.

Interestingly, several peptides that were negatively correlated with age became positively correlated upon methionine oxidation. This observation was consistent for two fragments of type I collagen alpha-1 as well as fragments of type IX collagen alpha-3, type XXV collagen alpha-1, sodium/potassium-transporting ATPase subunit gamma and retinol-binding protein 4 (Supplemental Table 1).

Age-correlation analysis in the diseased group

Next we studied the correlation of urinary peptides with ageing in the 10333 peptidomes of diseased individuals to determine potential discrepant and concerted correlations compared to healthy individuals. Individuals with pathological conditions were more likely to be older compared to healthy individuals (Table 1). Out of the 116 age-correlated peptides in healthy individuals, 112 were also found to correlate in diseased individuals. However, lower correlation coefficients were observed in the diseased compared to the healthy group (Supplemental Table 1). This observation was expected, given the assumed increased heterogeneity as a result

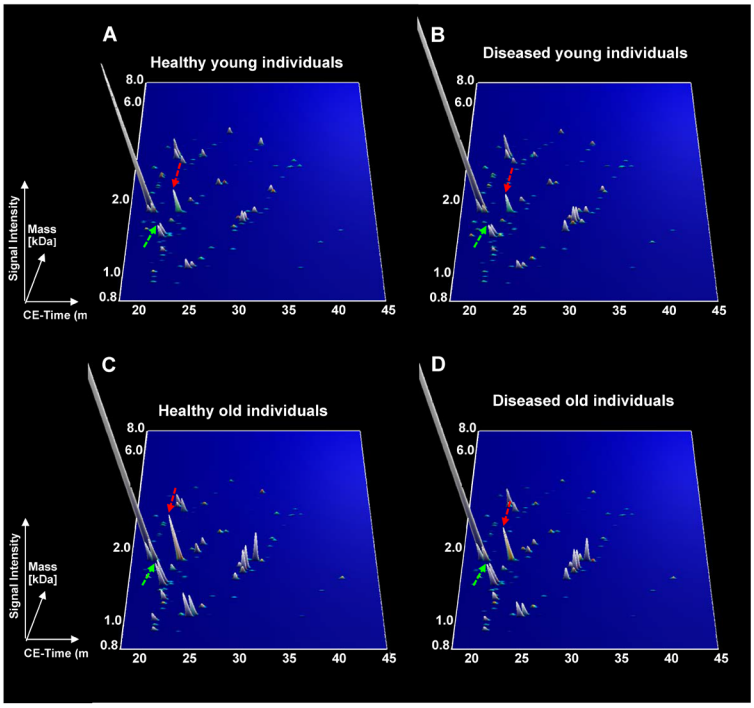

Figure 1: Urinary peptide marker pattern for the differentiation between healthy and diseased individuals. A. Healthy young between 20-29 years of age. **B.** Diseased young between 20-29 years of age. **C.** Healthy old from 60 years old of age and above. **D.** Diseased old from 60 years old of age and above. Only the mean intensity for each peptide was represented.

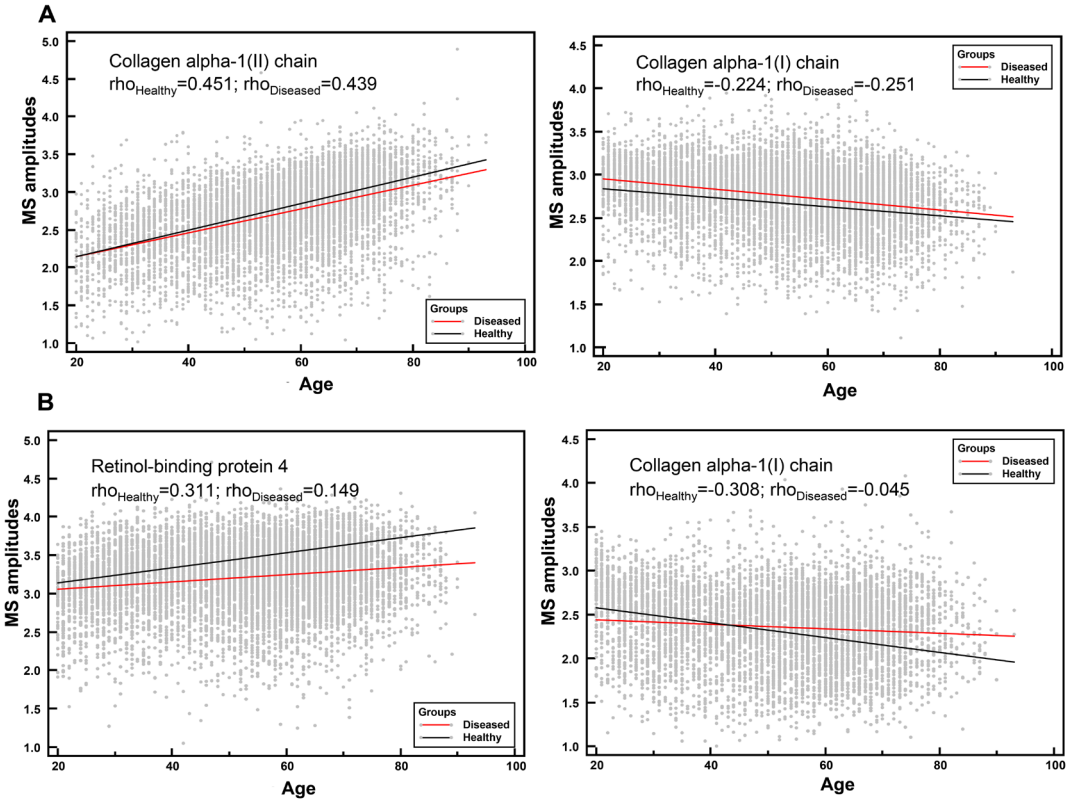

Figure 2: Correlation analysis of individual urinary peptides in healthy and diseased groups with age. A. Disease-unaffected peptides, collagen alpha-1(II) chain ($\rho_{Healthy}$ = 0.451, $p < 0.0001$ and $\rho_{Diseased}$ = 0.439, $p < 0.0001$) and collagen alpha-1(I) chain ($\rho_{Healthy}$ = -0.224, $p < 0.0001$ and $\rho_{Diseased}$ = -0.251, $p < 0.0001$). **B.** Disease-affected peptides, retinol-binding protein 4 ($\rho_{Healthy}$ = 0.311, $p < 0.0001$ and $\rho_{Diseased}$ = 0.149, $p < 0.0001$) and collagen alpha-1(I) chain ($\rho_{Healthy}$ = -0.308, $p < 0.0001$ and $\rho_{Diseased}$ = -0.045, $p < 0.0001$).

Table 1: Patient characteristics

	Healthy	Diseased
N (number of individuals)	1227	10333
Age (years)*	38.6 ± 12.4	54.4 ± 15.3
Sex (Male/Female)	623/604	6237/4096

* p-value <0.0001

Table 2: Different pathological conditions represented in the diseased group.

Diseases	N (number of individuals)
Alzheimer's	134
Bladder cancer	286
Cardiovascular diseases	1681
Diabetes mellitus	1715
Virus-triggered diseases (e.g. hepatitis, HIV)	332
Hepatocellular carcinoma	40
Kidney diseases	2154
Kidney diseases (transplanted)	430
Leukaemia	1622
Obesity	218
Pancreatic cancer	51
Polycystic ovary syndrome	73
Pheochromocytoma	11
Pregnancy	278
Pathologies related to the prostate	1217
Renal carcinoma	91
Total	10333

of various underlying pathologies. The 4 peptides not confirmed in the diseased group comprised three collagen fragments and a fibrinogen alpha chain fragment and were not considered for further investigations.

To determine if the 112 peptides were able to distinguish between young and old individuals in both healthy and diseased groups, the abundance of these peptides was studied in a dichotomous analysis in subpopulations of young versus old (Figure 1). Proteome profiles of young compared to old healthy individuals presented more visual differences than the profiles of young compared to old diseased individuals. However, the 112 age-correlated peptides were still able to distinguish between young and old individuals in healthy and diseased individuals. Interestingly, some peptides showed similar mean amplitudes in the healthy and diseased groups including for instance collagen alpha-1 (XXV) chain (Figure 1, green arrows) whereas other peptides such as collagen alpha-1 (III) chain (Figure 1, red arrows) depicted different amplitude profiles. Differences in age-correlated peptides were further investigated by comparing the correlation coefficients of the 112 peptides in both groups. As a result, peptides could be arranged into two groups: disease-unaffected and disease-affected peptides. These were defined by a non-significant (disease-unaffected) and a significant (disease-affected) p-value in the comparison of correlation coefficients between healthy and diseased groups (Supplemental Table 1, column "healthy vs diseased"). For instance among the best correlated peptides, the correlation coefficients for collagen alpha-1 (II) chain in healthy (rho = 0.451, $p <$ 0.0001) and diseased (rho = 0.439, $p < 0.0001$) individuals did not differ significantly (p = 6.21E-01) (Supplemental Table 1). An example for a disease-affected peptide is a fragment of collagen alpha-3(IV) chain with correlation coefficients in healthy (rho = 0.504, $p < 0.0001$) and diseased (rho = 0.420, $p < 0.0001$) individuals that differed significantly ($p < 0.0001$) (Supplemental Table 1). Further disease-unaffected peptides comprised fragments of type II alpha-1 and type III alpha-1 collagen (Figure 2A), while fragments of retinol-binding protein 4 and type I collagen alpha-1 were further disease-affected peptides (Figure 2B). Overall, 27 peptides widely represented by collagen fragments (89%) were disease-unaffected, while disease-affected peptides totalled with 85 and only included 66% of collagen fragments (Figure 3).

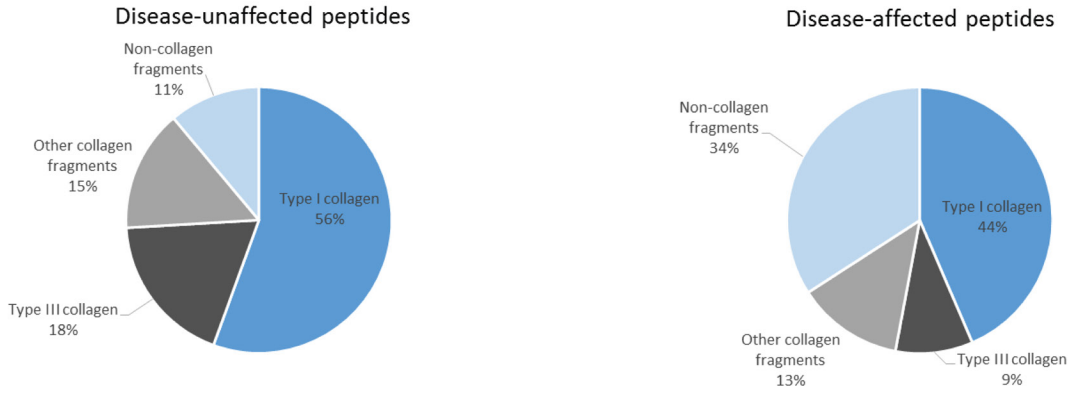

Figure 3: Comparison of age-correlated peptides identified in the healthy and diseased groups. A. Disease-unaffected peptides. **B.** Disease-affected peptides.

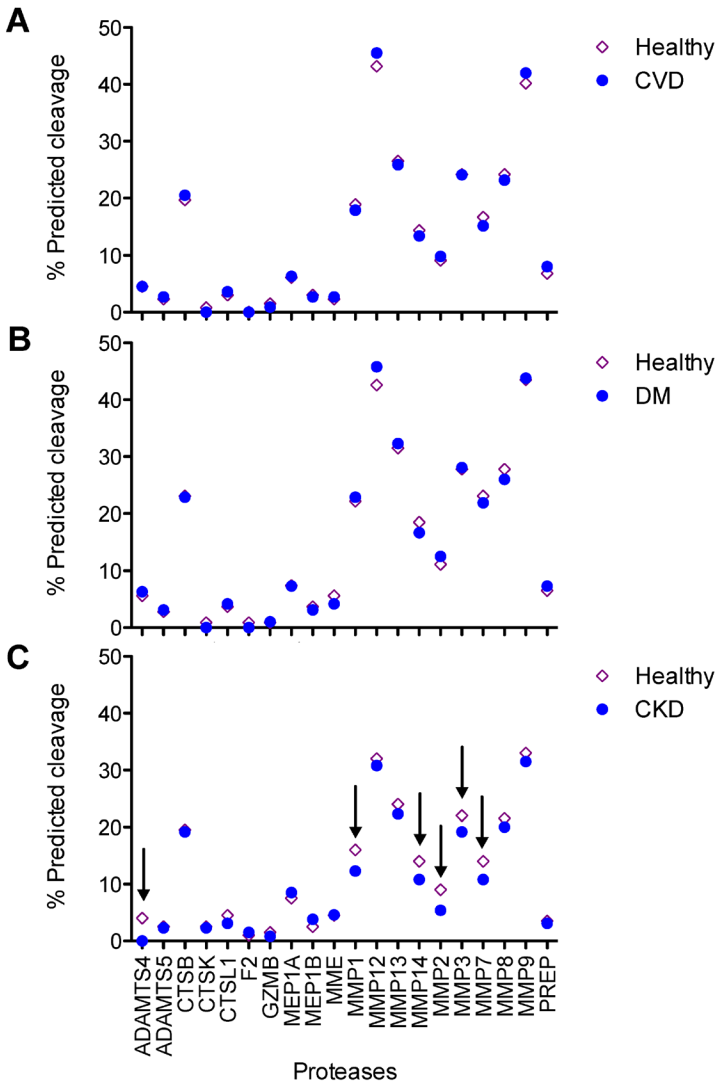

Figure 4: Comparison of age-correlated proteases between healthy individuals and disease subgroups. A. Cardiovascular diseases (CVD). **B.** Diabetes Mellitus (DM). **C.** Chronic kidney diseases (CKD). Arrows underscore the main changes in predicted protease activity between age-correlated disease-affected peptides in the healthy and the disease subgroups. ADAMTS4: A disintegrin and metalloproteinase with thrombospondin motifs 4; CTSB: cathepsin B; CTSK: cathepsin K; CTSL1: cathepsin L1; F2: thrombin; GZMB: granzyme B; MEP1A: meprin A subunit alpha; MEP1B: meprin A subunit beta; MME: neprilysin; PREP: prolyl endopeptidase.

Pathology-specific investigation of age-correlated peptides in different subgroups

As the diseased group of 10333 individuals included heterogeneous pathologies (Table 2), the pathology-specific age-association of the 112 peptides was investigated in three different more homogenous disease subgroups selected from the 10333 diseased individuals and then compared to the healthy group. The cardiovascular diseases (CVD, $n = 1681$) subgroup included individuals with heart failure, coronary artery disease and acute coronary syndrome. The chronic kidney diseases (CKD, $n = 2154$) subgroup included individuals with several kidney disorders such as vasculitis and glomerulopathies, whereas the diabetes mellitus (DM, $n = 1560$) subgroup consisted of type 1 and type 2 DM individuals with no detectable kidney disease.

The comparison of correlation analyses in all disease subgroups and the healthy group based on the 112 identified peptides provided an assessment of the distribution of age-correlated peptides. Sixty-six, 100 and 54 of the 112 age-correlated peptides were significantly correlated to age in individuals with CVD, CKD and DM, respectively (Supplemental Table 2). Thereby the number of disease-unaffected age-correlated peptides in the disease subgroups was reduced in CVD ($n = 10$) and DM ($n = 6$) compared to CKD ($n = 35$) (Supplemental Table 2). There was no overlap among these disease-unaffected peptides. As seen in the full diseased cohort of 10333 individuals, disease-affected non-collagen peptides were also almost two times as abundant as disease-unaffected ones in the three disease subgroups (Figure S1). In regard of collagen fragments, in the CVD subgroup 60% of disease-unaffected peptides originated from type I collagen compared to 46% of the disease-affected peptides. In comparison, type III collagen fragments comprised 30% of disease-unaffected peptides but only 12.5% disease-affected peptides. Type I collagen fragments in the CKD subgroup represented 60% of diseased-unaffected peptides compared to 43% in the disease-affected peptides (Supplemental Figure 1).

Prediction of protease activities

Based on the N- and C-terminal sequences of naturally occurring peptides, protease activity responsible for their generation can be predicted [5]. The *in silico* prediction of ageing-related changes in the activity of proteases potentially involved in the generation of the 112 peptides was based on the cleavage site consensus sequences of proteases and mean peptide intensities in individual healthy study subjects ($n = 1227$). The analysis resulted in 674 protease/cleavage associations related to 37 unique proteases. Amongst those, 32 proteases showed a significant correlation of their predicted activity with age in the healthy group (Supplemental Table 3). This comprised positive age-correlated activities e.g. of meprin A beta subunit, kallikrein 5, and thrombin as well as negative age-correlated activities e.g. of neprilysin, cathepsin L1, and matrix metalloproteinase-14 (MMP-14). We next compared predicted protease activities targeting disease-affected peptides between the healthy group and the disease subgroups (Figure 4). However, while we did not observe any significant differences between healthy individuals and individuals with CVD or DM, differences in age-related activities of A disintegrin metalloproteinase with thrombospondin motifs 4 (Adamts4) and MMPs appeared to be present in individuals with CKD (Figure 4, arrows).

Pathway enrichment analysis

Reactome pathway analysis for the identified disease-unaffected or disease-affected age-correlated peptides combined with the predicted proteases (in gene symbols) using ClueGO and CluePedia software resulted in several molecular pathways being significantly affected in the context of ageing (Figure 5, Supplemental Table 4). For disease-unaffected pathway analysis, the 27 peptides were combined with 19 predicted proteases that generated these peptide sequences. The same approach was performed for the 85 disease-affected peptides with the 32 corresponding predicted proteases. The network illustrates each pathway as individual nodes, while edges between pathways denote an approximation of biological interaction between the pathways based on the cross-pathway feature overlap. The analysis using disease-unaffected peptides revealed 6 molecular pathways associated with ageing including degradation of the extracellular matrix (ECM), activation of matrix metalloproteinases, collagen degradation, assembly of collagen fibrils, trafficking and processing of endosomal Toll-like receptors (TLRs) and endosomal/vacuolar pathway being enriched. However the analysis using disease-affected peptides, these six pathways were confirmed, and, in addition, degradation of IGF binding proteins was enriched in addition to the other 6 pathways (Figure 5).

DISCUSSION

The urinary proteome profiles of a unique cohort of 11560 individuals with an age ranging from 20 to 86 years were analysed with the aim of detecting specific ageing-associated urinary peptides and thus expand the current knowledge on the protein level and investigate the proteomic transition from normal ageing to age-related pathological complications.

The most prominent finding of the study was that increased age is associated with a decrease in the urinary

excretion of fragments from collagens forming the fibrillar structure of the ECM, including type I, II, III and V [6]. This finding is consistent with a study reported by Zürbig et al. (2009) [3]. While 49 fragments of mainly type I and type III collagen fragments were found to be significantly age-associated, only 15 of these peptides including fibrinogen chain and several collagen fragments of the Zürbig et al. study were sequenced and fulfilled the quality criteria of the current study. A decrease in type I and type III collagens was also observed in a study evaluating the effect of ageing on skin in a Caucasian female population [7]. Although the study cohort comprised only 218 healthy women between 33 and 77 years of age, findings may well be extrapolated to a male population. A decrease in fibrillar collagens observed in this study may result from several processes including but not limited to impaired collagen synthesis and/or impaired degradation causing aberrant ECM remodelling [8]. However, while the urinary excretion of the majority of fibrillar collagen fragments identified in the current study decreased with age, we also observed an age-associated increase in the urinary excretion of a few specific fibrillar collagen fragments with increasing age, especially of type II and III. In regards to type II collagen, these findings may indicate on one hand tissue and organ-dependent differences in homeostasis since type II collagen is mainly present in cartilage [9] and on the other hand the increased likelihood to develop osteoarthritis with advanced age as it has been shown that urinary levels of type II collagen fragments increased with osteoarthritis [10]. The increased excretion of some type III collagen fragments could be attributed to homeostasis as type III collagen plays an important role in type I fibrillogenesis and cardiovascular development [11]. Furthermore, all collagen fragments containing oxidised methionine were positively correlated with age. This is a novel finding that indicates an accumulation of oxidative modifications associated with age, which may lead to increased degradation. Of note: the corresponding

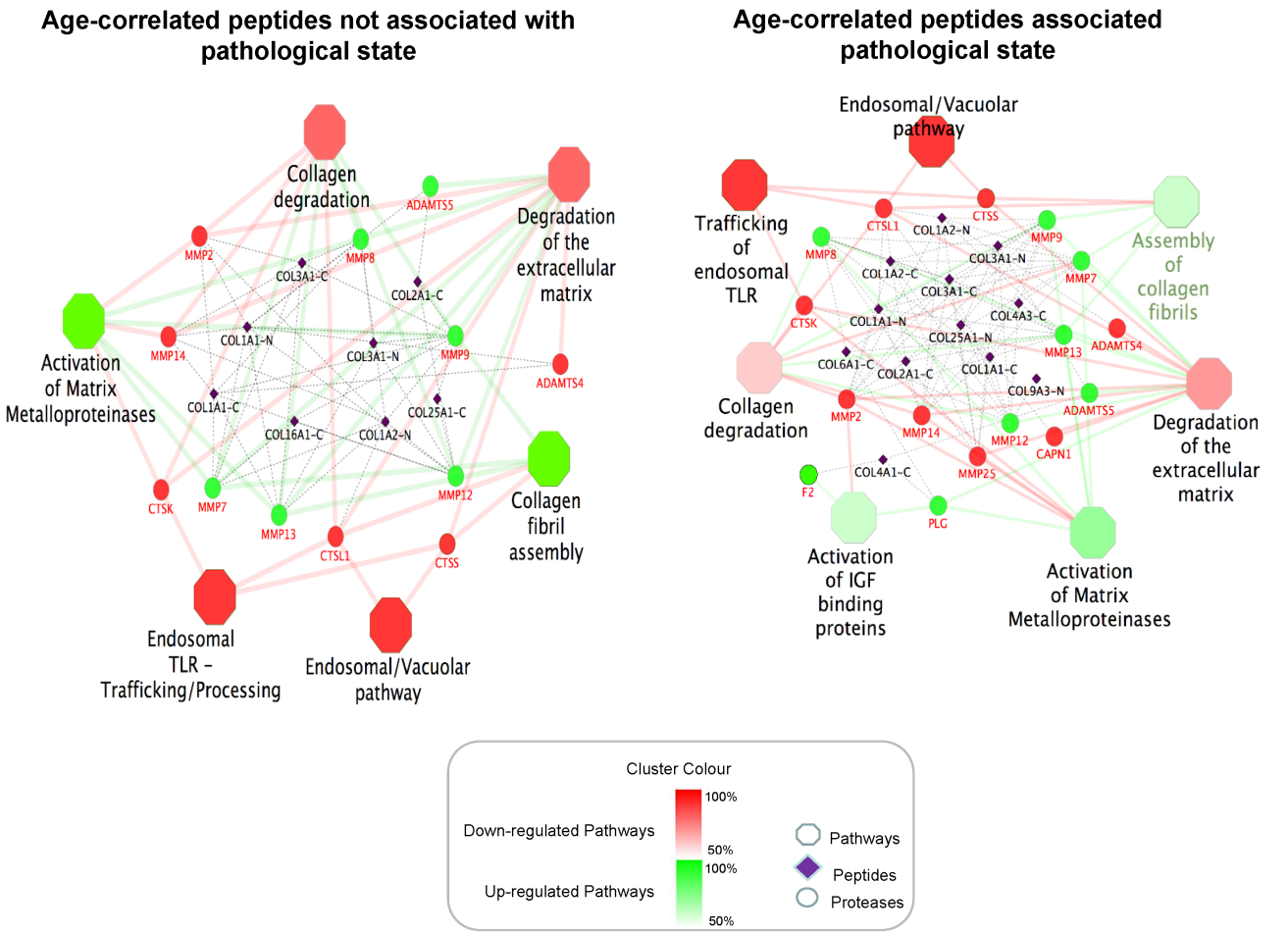

Figure 5: Molecular pathways associated with ageing. The network represents each pathway as individual octagonal node, while the circled nodes represent the predicted proteases that were targeted from the identified urinary peptides denoted in purple diamond nodes. The edges (links) between pathways denote an approximation of biological interaction between the pathways based on the cross-pathway feature overlap. Legends for the diamond nodes with a suffix of "-C/N" represent the peptide's cleavage site; i.e. "-C" for C-terminus and "-N" for the N-terminus."

unmodified peptide showed a decrease in urinary abundance with age. These findings may potentially also reflect progressive loss of control of oxidative stress during advancing ageing [12].

In addition to fibrillar collagens, the excretion of peptides from basement membrane collagens including type IV collagen alpha-1 (COL4A1) and alpha-3 (COL4A3) chains was also found to be altered. The excretion of a COL4A1-derived peptide was decreased and that of a peptide derived from COL4A3 was increased in advanced age. These findings are in agreement with the literature since COL4A1 was commonly found in the glomerular basement membrane of younger individuals whereas COL4A3 appears to be more common in adult individuals [13]. Increased urinary excretion of type IV collagen has furthermore been associated with renal dysfunction in patients with type 2 diabetes mellitus [14]. Hence, alterations of the basement membrane, readily observed in urine, are an important molecular event observed in ageing and renal impairment.

Of the non-collagenous peptides associated to ageing, fragments of uromodulin, beta-2-microglobulin and fibrinogen alpha and beta chains were most prevalent. The urinary excretion of most of these peptides showed a positive correlation with age. Fibrinogen, a glycoprotein involved in inflammation, and uromodulin, a kidney-specific protein, were shown to participate in renal fibrosis [15, 16]. The protein beta-2-microglobulin (B2M) is expressed in all nucleated cells and part of the light chain subunit of the major histocompatibility complex class I molecules [17]. Plasma and serum elevations of B2M were found to be associated with a plethora of pathological conditions including renal diseases [18] and cardiovascular diseases [19]. These peptides, that are in many cases also significantly associated with CKD [20], may reflect the reduction in kidney function observed in ageing [21].

The comparison of age correlation coefficients between healthy and diseased individuals enabled us to distinguish between disease-unaffected peptides reflecting processes of normal or healthy ageing and disease-affected peptides indicating events of pathological ageing. A correlation analysis in the diseased subgroups revealed greater similarity in age-correlated peptide excretions in urine between normal ageing and CKD compared to CVD and DM. This may indicate that urine does reflect the "status" of the kidney to a large degree [22]. Potential similarities in molecular alterations were suggested by decreases in excreted fibrillar collagen fragments, which is often indicative of alterations in the ECM turnover in the diseased kidney eventually resulting in fibrosis [23]. Furthermore, the protease analysis revealed a greater influence of CKD on age compared to CVD and DM. Our findings show that fibrosis developing in advanced age and CKD are similar whereas the similarity is less pronounced with fibrosis developing in CVD and DM.

A pathway enrichment analysis incorporating the 112 identified peptides and 32 predicted proteases suggested molecular pathways that are affected in normal and pathological ageing. Processes affected in normal ageing included perturbations in the collagen homeostasis, trafficking of toll-like receptors (TLRs) and endosomal pathways. As expected based on the abundance of collagen fragments, most of the molecular pathways found to be affected during ageing were involved in collagen homeostasis. Findings suggested accumulation of ECM or formation of fibrosis during ageing caused by a decrease in ECM degradation and an increase formation of collagen fibrils. These events result in a decrease of collagen fragments in the urine. Fibrosis is observed in renal ageing progressively degrading kidney function which potentially results in CKD [23]. Fibrosis in the heart can cause ventricular stiffening and impairment of heart function leading to cardiovascular diseases [24]. The enrichment analysis also indicated an impaired processing and trafficking of TLRs based-on the predicted negative age-correlation of the activities of cathepsin K, L1 and S. TLRs recognise molecular patterns that are broadly shared by pathogens and are essential for innate immune response by releasing cytokines and chemokines [25]. It was reported that cleavage of TLRs by cathepsins is crucial for the activation of TLRs signalling [26]. Therefore, attenuated cleavage of TLRs can contribute to perturbations in immunity in advanced age. Furthermore, cathepsins are endosomal proteases participating in diverse cellular processes including apoptosis, autophagy and necrosis [27]. Hence perturbations of the endosomal pathway should be more investigated in ageing.

Besides, the pathway analysis also enabled the identification of processes affected in pathological ageing. In addition to the molecular mechanisms affected in normal ageing, pathway enrichment analysis suggested an elevated degradation of insulin-like growth factor (IGF)-binding proteins (IGFBPs). The increased degradation of IGFBPs was predicted based on the activities of plasmin, thrombin and matrix metalloproteinase-12. Interactions between IGFBPs and IGFs generally have inhibitory effects on IGF-dependent signalling pathways potentially leading to augmented oxidative stress and inhibition of cellular proliferation, cellular differentiation and apoptosis [28, 29]. Interactions observed between ECM proteins, proteases including plasmin and thrombin [30] and IGFBPs contribute to the regulation of the bioavailability of IGFs [28]. Furthermore, increase in IGFBPs have been reported in patients with severe kidney failure [31] suggesting the involvement of the IGF pathway in pathology. IGF-1 has indeed been showed to decrease during ageing [32]. Thus, the activation of IGFBPs may be an important molecular event in ageing and further investigations are well justified to elucidate interactions between IGFBPs, plamin, thrombin and the ECM.

In conclusion, urinary proteome analysis enabled the

detection of ageing-associated peptides thereby generating considerable information about molecular pathways associated with normal ageing and pathological ageing. Perturbations in collagen homeostasis and trafficking of TLRs and endosomal pathways were generally observed in both normal and pathological ageing. However, increased degradation of the IGFBPs was additionally identified for the first time in ageing using urine samples. Besides, the comparison of urinary proteome profiles between healthy individuals and several diseased individuals revealed that protein fragments excreted in urine better depict similarities between normal ageing and CKD than CVD and DM. Findings demonstrated that with the help of appropriate technologies, urine can be used as a powerful biological fluid in ageing research.

MATERIALS AND METHODS

Ethics statement

The study was designed and conducted fulfilling all of the requisites of the laws on the protection of individuals collaborating in medical research and was in accordance with the principles of the Declaration of Helsinki.

Patient characteristics and CE-MS analysis

Patient data were retrieved from the "Human urinary database" dedicated to naturally occurring urinary proteins and peptides [4, 33]. All datasets included in the study were from previous studies, and all data were anonymised. The approach, employing anonymised proteomics data from previous studies, was approved by the local ethics committee. Datasets from 11560 individuals between 20 and 86 years of age were extracted (Table 1). The present cohort was divided into two groups: healthy and diseased. The healthy group included 1227 individuals and the diseased group 10333 individuals predominantly suffering from diabetes, cardiovascular and renal diseases (Table 2).

Sample preparation and capillary electrophoresis coupled to mass spectrometry analysis

For proteomic analysis, a 0.7 mL aliquot of urine was thawed immediately before use and diluted with 0.7 mL of 2 M urea, 10 mM NH_4OH containing 0.02% SDS. To remove higher molecular mass proteins, such as albumin and immunoglobulin G, the sample was ultra-filtered using Centrisart ultracentrifugation filter devices (20 kDa MWCO; Sartorius, Goettingen, Germany) at 3,000 rcf until 1.1 ml of filtrate was obtained. This filtrate was then applied onto a PD-10 desalting column (GE Healthcare, Uppsala, Sweden) equilibrated in 0.01% NH_4OH in HPLC-grade in H_2O (Roth, Germany) to decrease matrix effects by removing urea, electrolytes, salts, and to enrich polypeptides present. Finally, all samples were lyophilised, stored at 4°C, and suspended in HPLC-grade H_2O shortly before capillary electrophoresis coupled to mass spectrometry (CE-MS) analyses, as described [34].

CE-MS analyses were performed using a P/ACE MDQ capillary electrophoresis system (Beckman Coulter, Fullerton, USA) on-line coupled to a microTOF MS (Bruker Daltonics, Bremen, Germany) as described previously [34, 35]. The ESI sprayer (Agilent Technologies, Palo Alto, CA, USA) was grounded, and the ion spray interface potential was set between -4 and -4.5 kV. Data acquisition and MS acquisition methods were automatically controlled by the CE via contact-close-relays. Spectra were accumulated every 3 s, over a range of *m/z* 350 to 3000. Accuracy, precision, selectivity, sensitivity, reproducibility, and stability of the CE-MS measurements were demonstrated elsewhere [34].

Data processing

Mass spectral peaks representing identical molecules at different charge states were deconvoluted into single masses using MosaiquesVisu software. Only signals with $z>1$ observed in a minimum of 3 consecutive spectra with a signal-to-noise ratio of at least 4 were considered. Reference signals of 1770 urinary polypeptides were used for CE-time calibration by locally weighted regression. For normalisation of analytical and urine dilution variances, signal intensities were normalised relative to 29 ''housekeeping'' peptides [36]. The obtained peak lists characterise each polypeptide by its molecular mass [Da], normalised CE migration time [min] and normalised signal intensity. All detected peptides were deposited, matched, and annotated in a Microsoft SQL database allowing further statistical analysis [37]. For clustering, peptides in different samples were considered identical if mass deviation was < 50 ppm. CE migration time was controlled to be below 0.35 minutes after calibration.

Peptide sequencing

For sequencing of peptides the urine samples were analysed on a Dionex Ultimate 3000 RSLC nano flow system (Dionex, Camberly, UK) coupled to an Orbitrap Velos MS instrument (Thermo Fisher Scientific) as described in [38]. Data files were analysed using Proteome Discoverer 1.2 (activation type: HCD; min-max precursor mass: 790-6,000; precursor mass tolerance: 10 ppm; fragment mass tolerance: 0.05 Da; S/N threshold: 1) and were searched against the Uniprot human non-redundant database without enzyme specificity. No fixed

modifications were selected, oxidation of methionine, lysine and proline were selected as variable modifications. The peptide data were extracted using high confidence peptides, defined by an Xcorr ≥ 1.9, a delta mass between experimental and theoretical mass ± 5 ppm, absence of cysteines in the sequence (since cysteines without reduction and alkylation form disulphide bonds), absence of oxidised proline in protein precursors other than collagens or elastin, and top one peptide rank filters. For further validation of obtained peptide identifications, the strict correlation between peptide charge at the working pH of 2 and CE-migration time was used to prevent false identifications [39]. Only the sequenced peptides were further considered.

Correlation and statistical analyses

As peptide profiles across the samples were not normally distributed, we used the non-parametric Spearman's rank correlation coefficient for estimating the correlation of individual peptides using age as a continuous variable. All peptides present in the whole population were included in the correlation analysis since a frequency threshold was not set. The statistical significance was assumed at $p < 0.05$. The p-value was adjusted by applying Benjamini-Hochberg [40, 41]. A cut-off value was set for the correlation analysis and the coefficient of ≥ 0.2 or ≤ -0.2 ($|\rho| \geq 0.2$) was considered for further analysis. The analysis was performed using proprietary software (R-based statistic software, version 2.15.3) and verified with MedCalc version 8.2.1.0 (MedCalc Software, Mariakerke, Belgium). Graphs were generated using MedCalc. To discriminate between peptides affected by a diseased and those unaffected, the Spearman's rank correlation coefficient of a peptide was compared using MedCalc in healthy and diseased individuals.

In silico protease prediction

In order to link urinary fragments to the proteases involved, *in silico* protease mapping to urinary peptides was generated using Proteasix software as previously described [5]. Briefly, for each of the peptides, associated-proteases were predicted for both N and C-terminal cleavage sites. In parallel, a list of >6000 random octapeptide sequences was mapped using the same protocol in order to determine the specificity of the prediction. Only protease/cleavage site associations with higher prediction score than associations with random sequences were kept for further analysis. For each protease, predicted activity in each patient was calculated based on the mean of associated peptide intensities. A parametric Pearson correlation analysis between a predicted protease and the age was performed based on the mean intensities of cleaved peptides with the age of each individual.

Pathway enrichment analysis

For elucidating molecular pathways being associated to ageing, the age-correlated peptides and the significant proteases predicted by Proteasix were subjected to the Cytoscape's plug-ins ClueGo and CluePedia [42, 43]. Reactome pathway [44] served as the clustering criterion using a two-sided hypegeometry test followed by Bonferroni correction (significance level of 0.05) for identifying significantly affected pathways [45].

ACKNOWLEDGMENTS

The authors would like to thank Dr. Jochen Metzger, Dr. Adela Ramírez-Torres, Justyna Siwy and Maria Frantzi for suggestions for data analysis.

FUNDING

This work was supported by the grant GA 31354 (CodeAge) FP7-PEOPLE-2011-ITN.

CONFLICTS OF INTEREST

The authors have declared the following potential conflict of interest: H. Mischak is the founder and co-owner of Mosaiques Diagnostics GmbH, who developed the CE-MS technology for clinical application. E. Nkuipou-Kenfack, A. Bhat, M. Dakna, T. Koeck, and P. Zürbig are employees of Mosaiques Diagnostics GmbH.

REFERENCES

1. Lopez-Otin C, Blasco MA, Partridge L, Serrano M and Kroemer G. The hallmarks of aging. Cell. 2013;153:1194-1217.
2. Nkuipou-Kenfack E, Koeck T, Mischak H, Pich A, Schanstra JP, Zürbig P and Schumacher B. Proteome analysis in the assessment of ageing. Ageing Res Rev. 2014;18C:74-85.
3. Zürbig P, Decramer S, Dakna M, Jantos J, Good DM, Coon JJ, Bandin F, Mischak H, Bascands JL and Schanstra JP . The human urinary proteome reveals high similarity between kidney aging and chronic kidney disease. Proteomics. 2009;9:2108-2117.
4. Stalmach A, Albalat A, Mullen W and Mischak H. Recent advances in capillary electrophoresis coupled to mass spectrometry for clinical proteomic applications. Electrophoresis. 2013;34:1452-1464.
5. Klein J, Eales J, Zurbig P, Vlahou A, Mischak H and Stevens R. Proteasix: A tool for automated and large-scale

prediction of proteases involved in naturally occurring peptide generation. Proteomics. 2013;13:1077-1082.

6. Mouw JK, Ou G and Weaver VM. Extracellular matrix assembly: a multiscale deconstruction. Nat Rev Mol Cell Biol. 2014;15:771-785.
7. Baroni ER, Biondo-Simoes ML, Auersvald A, Auersvald LA, Montemor Netto MR, Ortolan MC and Kohler JN. Influence of aging on the quality of the skin of white women: the role of collagen. Acta Cir Bras. 2012;27:736-740.
8. Bonnans C, Chou J and Werb Z. Remodelling the extracellular matrix in development and disease. Nat Rev Mol Cell Biol. 2014;15:786-801.
9. Wilusz RE, Zauscher S and Guilak F. Micromechanical mapping of early osteoarthritic changes in the pericellular matrix of human articular cartilage. Osteoarthritis Cartilage. 2013;21:1895-1903.
10. Ishijima M, Kaneko H and Kaneko K. The evolving role of biomarkers for osteoarthritis. Ther Adv Musculoskelet Dis. 2014;6:144-153.
11. Liu X, Wu H, Byrne M, Krane S and Jaenisch R. Type III collagen is crucial for collagen I fibrillogenesis and for normal cardiovascular development. Proc Natl Acad Sci U S A. 1997;94:1852-1856.
12. Stadtman ER, Van RH, Richardson A, Wehr NB and Levine RL. Methionine oxidation and aging. Biochim Biophys Acta. 2005;1703:135-140.
13. Hudson BG, Tryggvason K, Sundaramoorthy M and Neilson EG. Alport's syndrome, Goodpasture's syndrome, and type IV collagen. N Engl J Med. 2003;348:2543-2556.
14. Araki S, Haneda M, Koya D, Isshiki K, Kume S, Sugimoto T, Kawai H, Nishio Y, Kashiwagi A, Uzu T and Maegawa H . Association between urinary type IV collagen level and deterioration of renal function in type 2 diabetic patients without overt proteinuria. Diabetes Care. 2010;33:1805-1810.
15. Docherty NG and Godson C. Fibrinogen as a damage-associated mitogenic signal for the renal fibroblast. Kidney Int. 2011;80:1014-1016.
16. Scolari F, Izzi C and Ghiggeri GM. Uromodulin: from monogenic to multifactorial diseases. Nephrol Dial Transplant. 2015;30:1250-1256.
17. Chiou SJ and Chen CH. Decipher beta2-microglobulin: gain- or loss-of-function (a mini-review). Med Sci Monit Basic Res. 2013;19:271-273.
18. Astor BC, Shafi T, Hoogeveen RC, Matsushita K, Ballantyne CM, Inker LA and Coresh J. Novel markers of kidney function as predictors of ESRD, cardiovascular disease, and mortality in the general population. Am J Kidney Dis. 2012;59:653-662.
19. Kals J, Zagura M, Serg M, Kampus P, Zilmer K, Unt E, Lieberg J, Eha J, Peetsalu A and Zilmer M. beta2-microglobulin, a novel biomarker of peripheral arterial disease, independently predicts aortic stiffness in these patients. Scand J Clin Lab Invest. 2011;71:257-263.
20. Schanstra JP, Zurbig P, Alkhalaf A, Argiles A, Bakker SJ, Beige J, Bilo HJ, Chatzikyrkou C, Dakna M, Dawson J, Delles C, Haller H, Haubitz M, et al. Diagnosis and Prediction of CKD Progression by Assessment of Urinary Peptides. J Am Soc Nephrol. 2015;26:1999-2010.
21. Zhou XJ, Rakheja D, Yu X, Saxena R, Vaziri ND and Silva FG. The aging kidney. Kidney Int. 2008;74:710-720.
22. Pejchinovski M, Hrnjez D, Ramirez-Torres A, Bitsika V, Mermelekas G, Vlahou A, Zürbig P, Mischak H, Metzger J and Koeck T. Capillary zone electrophoresis on-line coupled to mass spectrometry: a perspective application for clinical proteomics. Proteomics Clin Appl. 2015;9:453-468.
23. Anderson S, Halter JB, Hazzard WR, Himmelfarb J, Horne FM, Kaysen GA, Kusek JW, Nayfield SG, Schmader K, Tian Y, Ashworth JR, Clayton CP, Parker RP, et al. Prediction, progression, and outcomes of chronic kidney disease in older adults. J Am Soc Nephrol. 2009;20:1199-1209.
24. Biernacka A and Frangogiannis NG. Aging and Cardiac Fibrosis. Aging Dis. 2011;2:158-173.
25. Ma J, Chadban SJ, Zhao CY, Chen X, Kwan T, Panchapakesan U, Pollock CA and Wu H. TLR4 activation promotes podocyte injury and interstitial fibrosis in diabetic nephropathy. PLoS ONE. 2014;9:e97985.
26. Garcia-Cattaneo A, Gobert FX, Muller M, Toscano F, Flores M, Lescure A, Del Nery E and Benaroch P. Cleavage of Toll-like receptor 3 by cathepsins B and H is essential for signaling. Proc Natl Acad Sci U S A. 2012;109:9053-9058.
27. Turk B and Turk V. Lysosomes as "suicide bags" in cell death: myth or reality? J Biol Chem. 2009;284:21783-21787.
28. Kim KS, Seu YB, Baek SH, Kim MJ, Kim KJ, Kim JH and Kim JR. Induction of cellular senescence by insulin-like growth factor binding protein-5 through a p53-dependent mechanism. Mol Biol Cell. 2007;18:4543-4552.
29. Gong Z, Kennedy O, Sun H, Wu Y, Williams GA, Klein L, Cardoso L, Matheny RW, Hubbard GB, Ikeno Y, Farrar RP, Schaffler MB, Adamo ML, et al. Reductions in serum IGF-1 during aging impair health span. Aging Cell. 2014;13:408-418.
30. Booth BA, Boes M, Dake BL, Knudtson KL and Bar RS. IGFBP-3 binding to endothelial cells inhibits plasmin and thrombin proteolysis. Am J Physiol Endocrinol Metab. 2002;282:E52-E58.
31. Ivarsen P, Chen JW, Tietze I, Christiansen JS, Flyvbjerg A and Frystyk J. Marked reductions in bioactive insulin-like growth factor I (IGF-I) during hemodialysis. Growth Horm IGF Res. 2010;20:156-161.
32. Higashi Y, Sukhanov S, Anwar A, Shai SY and Delafontaine P. Aging, atherosclerosis, and IGF-1. J Gerontol A Biol Sci Med Sci. 2012;67:626-639.
33. Siwy J, Mullen W, Golovko I, Franke J and Zürbig P. Human urinary peptide database for multiple disease

biomarker discovery. Proteomics Clin Appl. 2011;5:367-374.

34. Mischak H and Schanstra JP. CE-MS in biomarker discovery, validation, and clinical application. Proteomics Clin Appl. 2011;5:9-23.
35. Arab S, Gramolini AO, Ping P, Kislinger T, Stanley B, van Eyk J, Ouzounian M, MacLennan DH, Emili A and Liu PP. Cardiovascular proteomics: tools to develop novel biomarkers and potential applications. J Am Coll Cardiol. 2006;48:1733-1741.
36. Jantos-Siwy J, Schiffer E, Brand K, Schumann G, Rossing K, Delles C, Mischak H and Metzger J . Quantitative Urinary Proteome Analysis for Biomarker Evaluation in Chronic Kidney Disease. J Proteome Res. 2009;8:268-281.
37. Gu YM, Thijs L, Liu YP, Zhang Z, Jacobs L, Koeck T, Zürbig P, Lichtinghagen R, Brand K, Kuznetsova T, Olivi L, Verhamme P, Delles C, et al. The urinary proteome as correlate and predictor of renal function in a population study. Nephrol Dial Transplant. 2014;29:2260-2268.
38. Kuznetsova T, Mischak H, Mullen W and Staessen JA. Urinary proteome analysis in hypertensive patients with left ventricular diastolic dysfunction. Eur Heart J. 2012;33:2342-2350.
39. Zürbig P, Renfrow M.B., Schiffer E, Novak J, Walden M, Wittke S, Just I, Pelzing M, Neusüss C, Theodorescu D, Root KE, Ross MM and Mischak H. Biomarker discovery by CE-MS enables sequence analysis via MS/MS with platform-independent separation. Electrophoresis. 2006;27:2111-2125.
40. Benjamini Y and Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Royal Stat Soc B (Methodological). 1995;57:125-133.
41. Benjamini Y and Yekutieli D. The control of the false discovery rate under dependency. Ann Stat. 2001;29:1165-1188.
42. Bindea G, Mlecnik B, Hackl H, Charoentong P, Tosolini M, Kirilovsky A, Fridman WH, Pagès F, Trajanoski Z and Galon J. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics. 2009;25:1091-1093.
43. Bindea G, Galon J and Mlecnik B. CluePedia Cytoscape plugin: pathway insights using integrated experimental and in silico data. Bioinformatics. 2013;29:661-663.
44. Haw R, Hermjakob H, D'Eustachio P and Stein L. Reactome pathway analysis to enrich biological discovery in proteomics data sets. Proteomics. 2011;11:3598-3613.
45. Saito M, Kimoto M, Araki T, Shimada Y, Fujii R, Oofusa K, Hide M, Usui T, Yoshizato K. Proteome analysis of gelatin-bound urinary proteins from patients with bladder cancers. Eur Urol. 2005;48:865-871.