

MANTI: Automated Annotation of Protein N-Termini for Rapid Interpretation of N-Terminome Data Sets

Fatih Demir,* Jayachandran N. Kizhakkedathu, Markus M. Rinschen, and Pitter F. Huesgen*

Cite This: *Anal. Chem.* 2021, 93, 5596–5605

Read Online

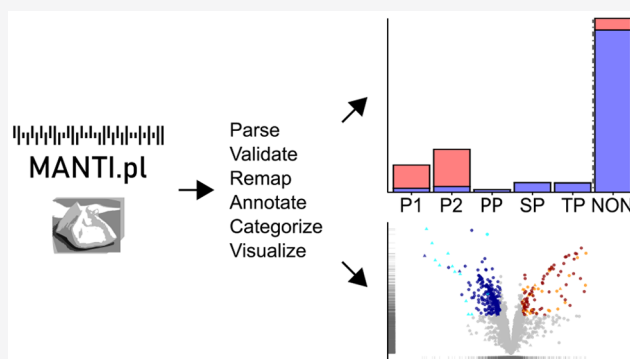
ACCESS |

Metrics & More

Article Recommendations

Supporting Information

ABSTRACT: Site-specific proteolytic processing is an important, irreversible post-translational protein modification with implications in many diseases. Enrichment of protein N-terminal peptides followed by mass spectrometry-based identification and quantification enables proteome-wide characterization of proteolytic processes and protease substrates but is challenged by the lack of specific annotation tools. A common problem is, for example, ambiguous matches of identified peptides to multiple protein entries in the databases used for identification. We developed MaxQuant Advanced N-termini Interpreter (MANTI), a stand-alone Perl software with an optional graphical user interface that validates and annotates N-terminal peptides identified by database searches with the popular MaxQuant software package by integrating information from multiple data sources. MANTI utilizes diverse annotation information in a multistep decision process to assign a conservative preferred protein entry for each N-terminal peptide, enabling automated classification according to the likely origin and determines significant changes in N-terminal peptide abundance. Auxiliary R scripts included in the software package summarize and visualize key aspects of the data. To showcase the utility of MANTI, we generated two large-scale TAILS N-terminome data sets from two different animal models of chemically and genetically induced kidney disease, puromycin adenonucleoside-treated rats (PAN), and heterozygous Wilms Tumor protein 1 mice (WT1). MANTI enabled rapid validation and autonomous annotation of >10 000 identified terminal peptides, revealing novel proteolytic proteoforms in 905 and 644 proteins, respectively. Quantitative analysis indicated that proteolytic activities with similar sequence specificity are involved in the pathogenesis of kidney injury and proteinuria in both models, whereas coagulation processes and complement activation were specifically induced after chemical injury.



INTRODUCTION

Proteolysis is a regulatory protein modification that results in the generation of distinct proteoforms with altered or new functions.^{1,2} Misregulated proteolytic processing is implicated in many diseases, including amyloid β -precursor protein (ABPP) processing in Alzheimer's disease,³ cancer micro-environment remodeling,⁴ and podocyte damage in kidney disease.⁵ Characterization of protein termini reveals affected substrate proteins and allows the determination of the underlying protease–substrate relationships, which is a critical requirement to qualify specific proteases as drug targets or antitargets in a disease.⁶

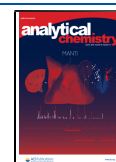
The standard “bottom-up” proteomics uses proteolytic digestion of proteins into the peptides for analysis by tandem mass spectrometry. The acquired peptide fragmentation spectra are then identified by computational matching to the predicted⁷ or previously collected⁸ peptide fragmentation spectra. Protein termini, including neotermini generated by endogenous proteolytic enzymes before proteome analysis, are only rarely identified in these standard procedures since they are only a minor fraction in the proteome digest, often of low

abundance, and typically not considered during the database matching due to their partial mismatch with the digestion enzyme specificity.^{2,9} To overcome these issues, a large variety of methods for the enrichment of protein termini have been developed (for reviews, see refs 1, 10, 11). Most regularly used are negative selection approaches, where free primary amines of native protein N-termini, including the protease-generated neo-N-termini, are chemically modified before enzymatic proteome digestion (Figure S1). All peptides generated by the enzymatic proteome digest carry a free primary amine, which is subsequently used for selective depletion, for example, by capture with an aldehyde-functionalized polymer in Terminal Amine Isotopic Labeling of Substrates (TAILS)¹²

Received: January 22, 2021

Accepted: March 4, 2021

Published: March 17, 2021



or for chemical modification and subsequent selection by altered chromatographic mobility in Charge-based/Combined FRActional DIagonal Chromatography (ChaFRADIC¹³/CO-FRADIC¹⁴). Recent developments have further simplified and increased the sensitivity of the N-termini enrichment procedures,^{15,16} enabled automation for high-throughput analysis,¹⁷ and improved the identification of specific subsets of the N-terminal peptides.^{18–20}

Different protein N-termini can arise from the translation of predicted and alternative initiation sites or spliced mRNA, known or predicted proteolytic maturation such as signal/transit peptide processing or propeptide cleavage, as well as unpredicted proteolytic activity.^{1,2} Correct positional annotation is therefore the basis for classification and, together with quantitative evaluation across experimental conditions, required for biological interpretation of N-terminome data and identification of protease substrates. However, this classification is challenging, as many identified N-terminal peptide sequences match multiple, often distinct, protein sequences in the database. Several web-based software tools visualize peptide position in relation to protein sequences and annotated features: Protter²¹ maps peptides to topological and domain features retrieved from UniProt,²² while QARIP²³ and Proteator²⁴ additionally visualize peptide abundance data and thereby reveal differentially affected domains, as for example, membrane protein shedding, in quantitative proteome data sets. All of these tools can assist in mapping identified protein termini to sequence features but generate one graphical depiction per protein that requires labor-intensive manual assessment and interpretation. Automated pipelines, on the other hand, rarely exist. CLIPPER²⁵ is the only previous tool specifically designed for N-terminal peptide annotation that leverages “razor” protein assignments from protein-level data from the same experiment according to well-established criteria.²⁶ However, CLIPPER is tightly integrated within the Trans-Proteomic Pipeline (TPP)²⁷ and not compatible with other pipelines. In summary, a one-stop-shop solution for the annotation of data generated by the plethora of different N-termini enrichment workflows does not exist yet.

Here we present MANTI, a new termini-centric software tool that facilitates validation, annotation, and quantitative evaluation of N-terminome data sets analyzed with MaxQuant.²⁸ MaxQuant is currently one of the most widely used software packages due to its ease of use and superior accuracy in peptide quantification, particularly with stable isotope labels. The key advantage of MANTI is the integration of an array of external annotation sources into an automated, hierarchical N-termini validation and annotation workflow that generates curated N-termini information (Figure 1). MANTI thereby provides a convenient, multidimensional overview over the data set, enables rapid identification of significant changes in substrates cleavage and links this with the annotation information that points to their potential physiological relevance. MANTI has already been successfully used for a multitude of different samples, including N-terminome data sets from both animals^{5,29} and plants.^{16,20,30,31}

We further demonstrate the utility of MANTI for identification, annotation, and quantitative evaluation of proteolytic cleavages in two large-scale data sets comprising >10 000 peptide identifications from two independent animal models of chronic kidney disease. With the help of the annotation workflow, we observed that both chemically

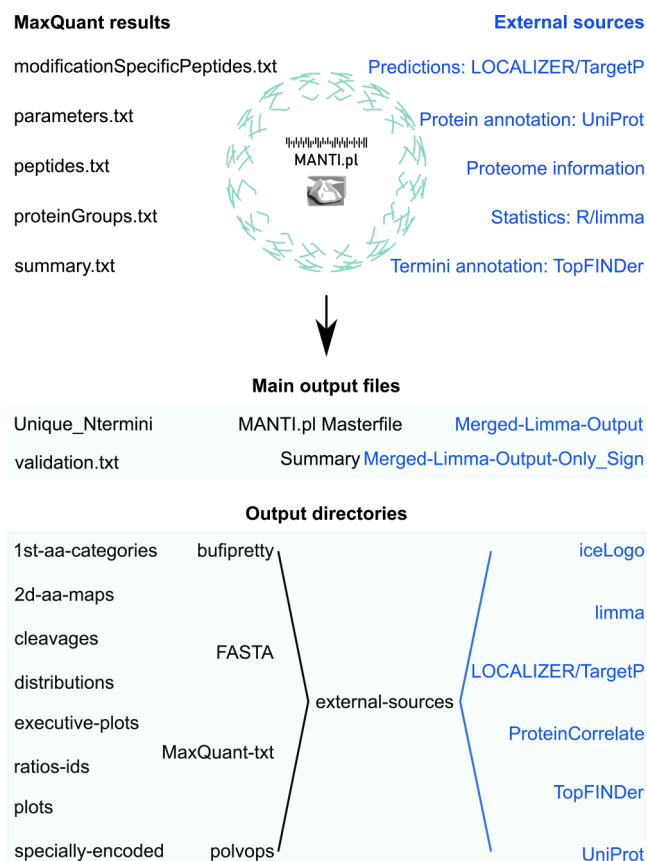


Figure 1. MaxQuant source files, MANTI output, and external sources, which can be optionally integrated (blue). MANTI can integrate external sources for validation and interpretation of the N-termini data: (1) annotations for the validated N-terminal peptides can be retrieved from UniProt; (2) statistical validation with the limma package for R; (3) targeting signal/transit peptide predictions from LOCALIZER and/or TargetP2.0; (4) associated protein identifications from proteome, e.g., aliquots before enrichment, from (5) TopFINDER annotation. The main output files are the MANTI Masterfile and summary; the validation.txt file allows monitoring of the MANTI filtering steps during the N-termini validation. N-terminal peptides are collapsed into unique N-termini by their first eight amino acids and grouped by protein ID in the Unique_N-termini file. Integrated external annotation sources are all copied into the “external-sources” folder for reference.

induced kidney injury in puromycin adenonucleoside (PAN)-treated rats^{32,33} and genetically caused kidney disease in heterozygous Wilms Tumor protein 1 (WT1)-deficient mice³⁴ are associated with the altered abundance of specific subsets of proteolysis-derived protein N-termini.

EXPERIMENTAL SECTION

Animal Material. Glomeruli from PAN- or control-treated rats and kidney cortices of *wf1*^{+/-} or control littermate mice were obtained from surplus material from our previous study.⁵ Protocols for animal experiments were approved by the applicable local regulatory bodies, the Landesamt für Natur, Umwelt und Verbraucherschutz Nordrhein-Westfalen, and Regierungspräsidium Freiburg (G15-134). Four biological replicates were collected for the PAN model⁵ and three for the WT1 model.³⁴

N-Terminal Peptide Enrichment by TAILS. TAILS N-termini enrichment was performed as described in detail

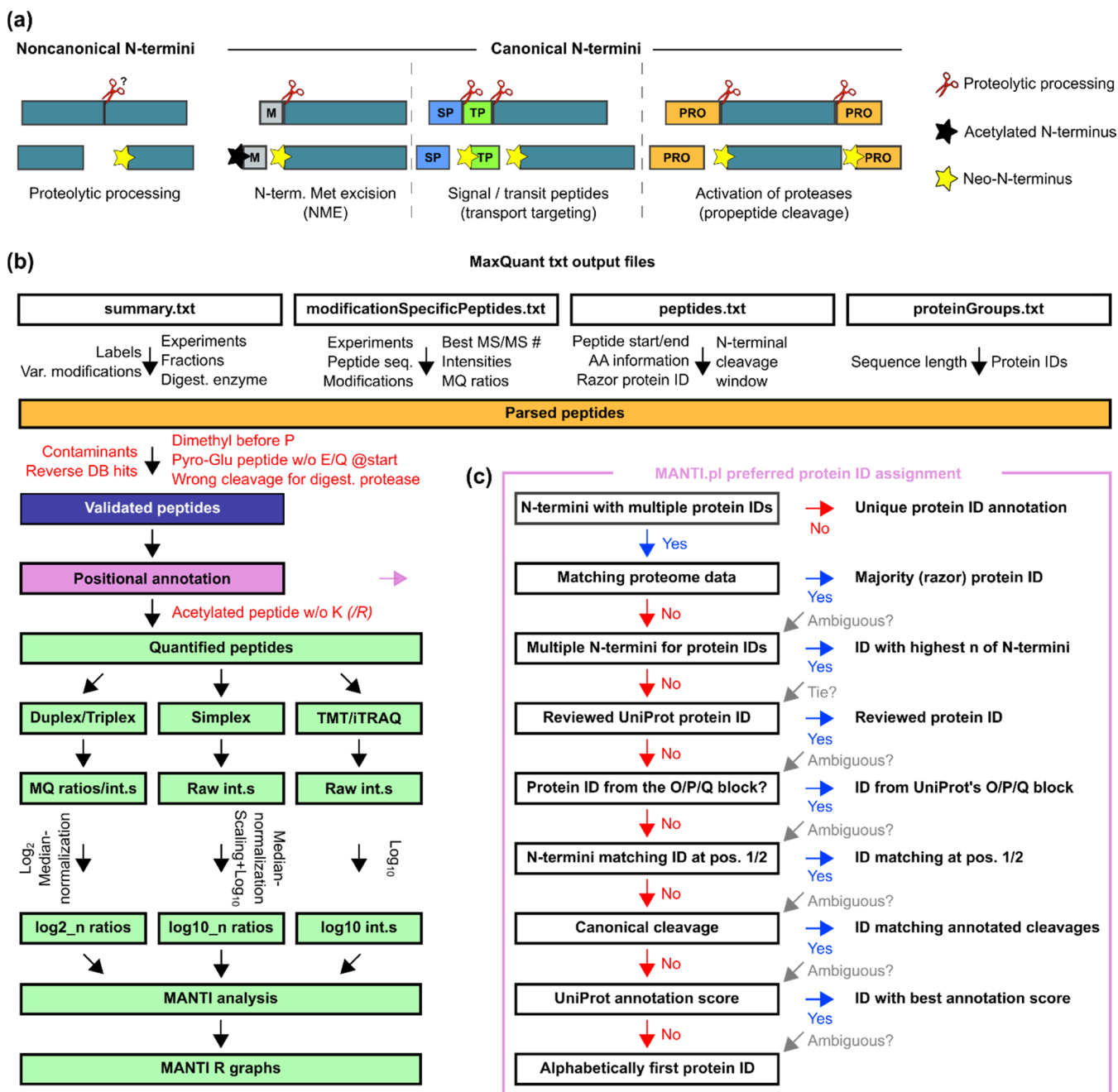


Figure 2. N-termini categorization and the MANTI workflow. (a) Protein N-termini can be categorized into canonical N-termini derived from protein synthesis with or without co-translational N-terminal methionine excision (NME), which removes the initiator-methionine, N-terminal signal and transit peptide cleavages for subcellular targeting, cleavage of known inhibitory propeptides converting zymogens into active proteases, and noncanonical N-termini generated by site-specific proteolytic processing. (b) Workflow of MANTI processing. MaxQuant output files (white) are parsed (orange), validated (purple), a preferred protein ID assigned (pink) and processed into quantitative MANTI analysis results (green). (c) Binary decision tree for the assignment of the preferred protein ID by MANTI during the positional annotation step.

elsewhere.⁵ In addition, enriched N-terminal peptides were fractionated at basic pH with steps of 15, 20, 25, 30, and 50% acetonitrile (ACN) in 10 mM ammonium hydroxide, followed by an acidic fraction with 50% ACN in 0.1% TFA using HR-X reverse-phase cartridges (Macherey-Nagel, Düren, Germany), subsequently purified by C₁₈ StageTips³⁵ and analyzed by nano-LC-MS/MS as described.⁵ The database queries were performed with MaxQuant v1.6.0.16 against the UniProt mouse (UniProt release 2017_10) and rat (2017_12) proteomes including isoform entries at a PSM FDR of 0.01 using standard settings for Bruker impactII Q-TOF instru-

ments. The enzyme specificity set to semispecific, free N-terminus ArgC as dimethylation prevents trypsin cleavage at Lys residues. Cys carbamidomethylation (+57.021464) was set as a fixed modification and MaxQuant labels were configured as light (+28.031300) and heavy dimethylation (+34.063117) on Lys residues and peptide N-termini, whereas Met Oxidation (+15.994914), peptide N-terminal acetylation (+42.010565), and N-terminal pyro-Glu formation from Glu (−18.010565) or Gln (−17.026549) were set as variable modifications.

MANTI Dependencies and Source Code. The only dependency for MANTI is a Perl interpreter, e.g., the widely

known Perl distributions ActiveState Perl (<https://www.activestate.com/products/perl>) or Strawberry Perl (<http://strawberryperl.com/>) that are available on a variety of operating systems; no additional modules or libraries are necessary.

MANTI is available from the project site at sourceforge.net as the source code or bundled with executable files for the Win10 and Linux platforms (<https://sourceforge.net/projects/manti/>); additionally, R packages are required for the production of the graphs that can be obtained via CRAN: GGally, UpSetR, ggcorrplot, ggpubr, ggrepel, stringr, and svglite. Currently, N-terminal modification by reductive dimethylation with a single formaldehyde isotope (simplex), two (duplex) or three (triplex) stable isotope dimethyl labels, and SILAC labeling and TMT/iTRAQ multiplex labeling are supported.

Statistical Analysis. The auxiliary `polvops.R` script integrated into MANTI performs a `limma`-moderated³⁶ *t*-test.

Data Visualization. MANTI contains two R³⁷ scripts that allow the optional output of a variety of data plots with the `ggplot2` library,³⁸ including volcano plots for statistical inference and UpSet plots³⁹ for data distribution among experiments. First, `polvops.R` incorporates statistical analysis and visualizes the results as sophisticated volcano plots, with the shape of the data points indicating the modification (circle, dimethylated; triangle, acetylated) and color hue reflecting the abundance change and classification (orange and red, significantly accumulating canonical and noncanonical termini; cyan and blue, depleted canonical and noncanonical termini) of the identified N-terminal peptides at the chosen cutoff values (here: LIMMA-moderated *t*-test *p*-value <0.05 and at least 50% change in abundance, $\log_2 \pm 0.58$). Those annotated volcano plots depicting the statistical significance and N-terminal categorization are generated for several, selected *p*-value and fold-change cutoff values. The second R script accompanying MANTI is called `bufipretty.R`, which generates a variety of data plots including a one-page PDF summary of the parsed N-terminome data.

Data Availability. Raw MS data have been deposited with the PRIDE repository⁴⁰ with the accession codes PXD023299 for the PAN data set and PXD023292 for the WT1het data set.

RESULTS AND DISCUSSION

MANTI is a Perl⁴¹ script that parses and validates peptide identifications from semispecific MaxQuant searches and integrates annotation from various sources to derive a consensus N-termini annotation to assist the N-terminome data interpretation. MANTI can be executed as a command-line Perl script or via the graphical user interface Yoğurtlu MANTI (Figure S2).

MANTI follows a peptide-centric approach and first retrieves experimental information on the digestion protease, experiment names, fraction numbers, variable modifications considered, and the applied labeling strategies from the `summary.txt` file of the MaxQuant search result output (Figure 2b). Subsequently, MANTI parses the peptide sequences, their modification (from the `modificationSpecificPeptides.txt`), and quantitative information such as unprocessed MS1 intensities and associated MaxQuant ratios or TMT/iTRAQ reporter ion intensities and retrieves associated information such as peptide positions, amino acid information, razor protein ID (from the `peptides.txt`), and protein IDs (from `proteinGroups.txt`). In the “validation” step, MANTI eliminates peptides marked as

contaminants or decoy database entries. N-terminal modifications considered during database queries typically include the amine modification introduced in the initial labeling step, such as dimethylation or TMT/iTRAQ, endogenous modifications such as peptide N-terminal acetylation, and pyro-Glutamate formation from N-terminal Gln/Glu, which can occur either as an endogenous modification or an artifact during the enrichment workflow. MANTI next removes nonsense identifications such as pyro-Glu-modified peptides lacking an N-terminal Glu or Gln, peptides with dimethyl label at the N-terminal proline residue (not possible as Pro is a secondary amine), and peptide identifications with mismatched last amino acids according to the protease specificity used for the semispecific MaxQuant database search. In addition to these validation filters, stable isotope-based quantification is tested for plausibility, for example, quantifications reported for acetylated peptides that do not contain any lysine in their sequence are removed. All validation steps performed by MANTI are reported and summarized in a `validation.txt` output file.

In the next step, validated N-terminal peptides are annotated with position-specific information from several web-based resources: Known N-terminal sequence features such as pro-, signal-, and transit peptide cleavage positions are retrieved online from UniProt,²² along with other sequence features such as transmembrane domains, motifs, and active sites (Figure 1). Termini-centric information such as known protease cleavage sites, alternative splice sites, and alternative translation start sites can be imported from TopFIND,⁴² and improved predictions of signal-or/transit peptides as well as chloroplast targeting sequences for plant data sets can be retrieved from LOCALIZER⁴³ or TargetP2.0.⁴⁴ Annotations and associated prediction scores are presented in the MANTI result, but also saved locally for future reference.

Sequences of N-terminal peptides, like any other single peptides, often match several protein entries in a proteome database, which may include closely related proteins with high sequence similarity, annotated isoforms, and redundant or outdated protein sequences not yet discarded by curation efforts. This ambiguity is a critical limitation that impedes further biological interpretation. To limit the number of hypotheses that need to be considered during the evaluation, MANTI utilizes all available sequence and annotation information in a binary decision tree to select a “preferred” UniProt protein accession number among all of the protein entries matching the identified peptide sequences (Figure 2b). If matching proteome data is available, the preferred razor protein determined by MaxQuant is retained. If not, or multiple protein entries remain, the number of N-terminal peptides matching to each protein ID drives the decision process. Adhering to Occam’s razor principle,⁴⁵ ambiguous peptides are counted for the protein ID with the highest number of peptides, and the protein ID with the highest number of N-terminal peptides in the data set is set as the preferred protein. If multiple protein entries remain, the existing annotations are utilized to determine conservative preferred entries: First, high-quality protein models with UniProt protein IDs marked as “reviewed” (manually curated) are preferred, followed by the well-annotated UniProt entries starting with O/P/Q, then protein IDs with cleavage at position 1/2 or an annotated canonical cleavage are preferred. Finally, if no preferred protein ID could be determined up to this point, the UniProt annotation score and as a last

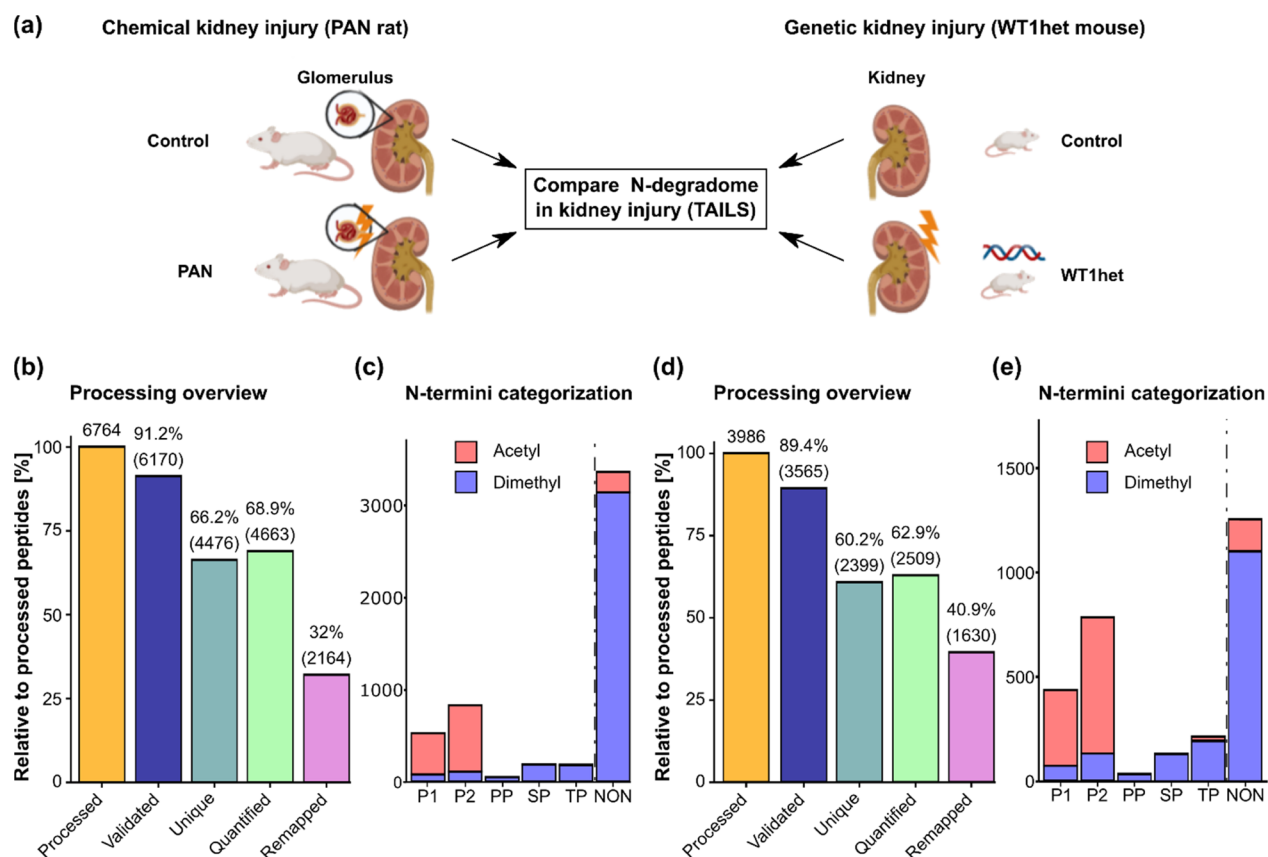


Figure 3. MANTI analysis of different kidney injury models. (a) Altered proteolytic processing in two models of chronic kidney injury, chemical treatment of rats with puromycin, and genetic deficiency in WT1 protein in mice were analyzed for altered proteolytic processing in comparison to the corresponding control. (b, c) MANTI evaluation of the PAN data set. (b) MANTI processing summary displaying the amount of MANTI-validated, unique, and quantified N-termini. Additionally, the number of reannotated preferred protein IDs is given. (c) N-termini categorization according to integrated annotation sources (UniProt, LOCALIZER/TargetP): canonical N-termini comprise mainly acetylated peptides, the mature protein N-terminus at position 1/2 or cleavages at signal (SP)/transit peptides (TP) and propeptides (PP). The remaining cleavages are mainly dimethylated and feature no expected cleavage, thus these cleavages are categorized as noncanonical (NON). (d, e) Respective graphs for the WT1 data set display a similar distribution of N-termini as the PAN data set with the majority of the identified N-termini being dimethylated and noncanonical.

consequence, the alphabetically first protein ID is selected as the preferred protein. MANTI thus places preference on high-quality protein models previously proven to exist, and on protein models indicating an origin by known synthesis and conserved processing events before postulating proteolytic processing events.

For data sets with multiple experimental conditions, MANTI integrates quantitative information either from label-free quantification or relative abundance ratios determined from stable isotope labeled experiments, filters incorrect quantifications in stable isotope labeling experiments (e.g., N-terminally acetylated peptides without lysine residues resulting in aberrantly high ratios in the light channel), and test for significant changes between the different conditions using a LIMMA-moderated *t*-test.³⁶ Finally, MANTI creates a summary of the validated peptides and the main output file listing all N-terminal peptide sequences, their modifications, and associated quantitative and annotation information in the main directory.

MANTI further exports several subfolders for detailed investigations of the N-terminal peptides and their properties, including frequencies of N-terminal amino acids and modifications in different subcellular compartments (Figure 1 and Table S1). Additionally, MANTI can automatically

generate a variety of data blots using two embedded R scripts to summarize and visualize the data.

Annotation of N-Terminal Peptides and Identification of Proteolytic Cleavages in Kidney Disease Models.

Proteases are emerging as key players in kidney disease.⁴⁶ We therefore chose to investigate altered processing in two different models of kidney disease, in rats treated with PAN as a chemical kidney injury model,⁵ and in heterozygous WT1 knockout mice (WT1het) as a genetic kidney injury model³⁴ to determine common and distinct changes in proteolytic processing patterns associated with podocyte injury (Figure 3a). We extracted proteins under denaturing conditions in the presence of protease inhibitors, differentially labeled primary amines in proteins from diseased animals and corresponding controls using different stable isotopologues of formaldehyde and enriched N-terminal peptides by negative selection using the well-established TAILS protocol.^{9,12} Subsequent MS/MS analysis and database searches with MaxQuant identified 6764 peptides in 4 biological replicates of PAN-treated rats and 3986 peptides in the WT1 mice, at an FDR of 1%. Nevertheless, the MANTI workflow flagged 8.8% of the parsed peptides in the PAN data set (Figure 3b) and 10.6% of the peptides in the WT1het data set (Figure 3d) as hits against the decoy database, peptides derived from common contaminants

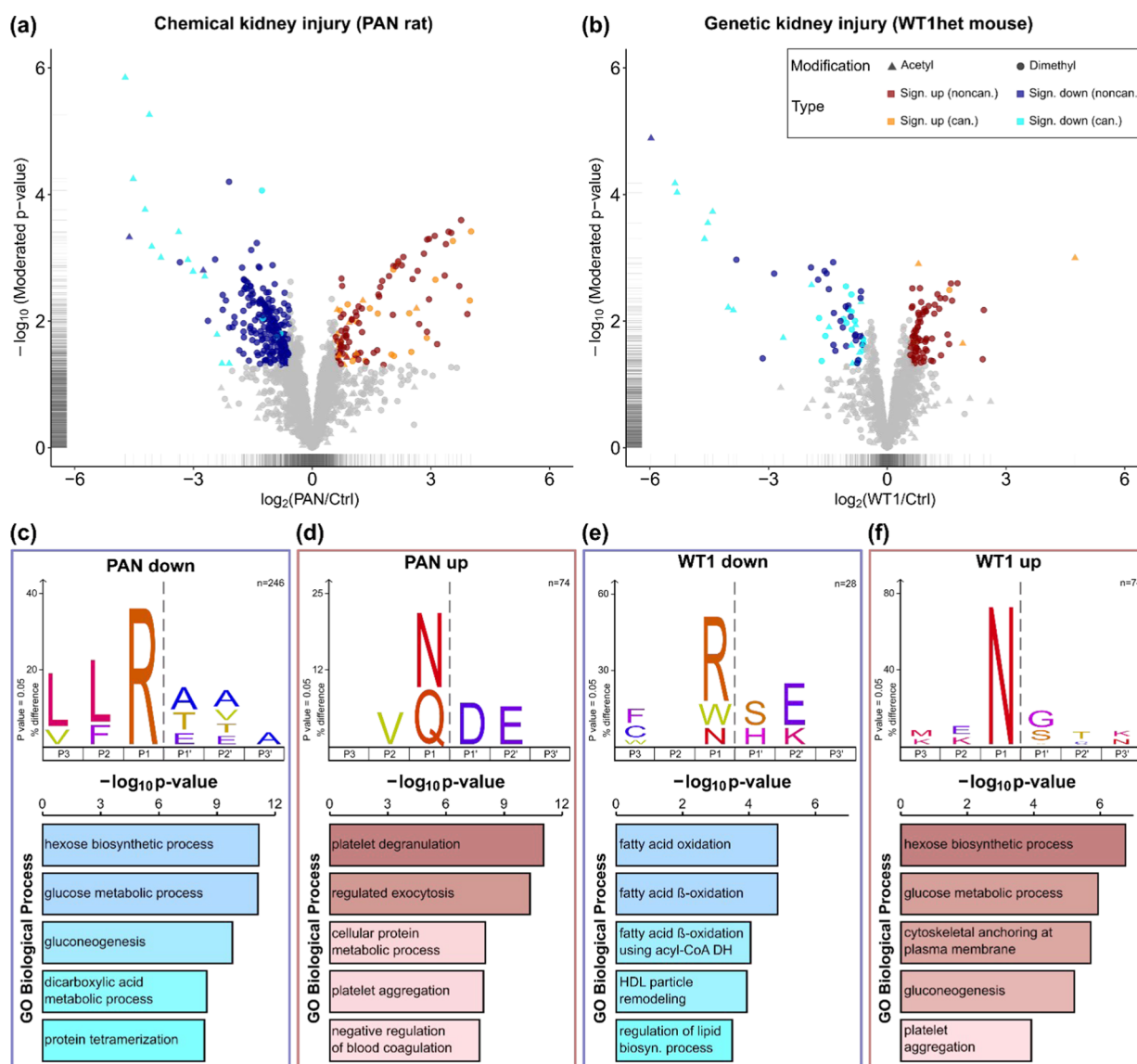


Figure 4. N-termini alterations in the PAN and WT1 kidney disease models. (a, b) Volcano plots generated with the polvops.R script display significantly accumulated ($\log_2 > 0.58$, LIMMA-moderated t -test $p < 0.05$) and depleted ($\log_2 < -0.58$, LIMMA-moderated t -test $p < 0.05$) canonical (orange/cyan) and noncanonical (red/blue) N-terminal peptides. N-terminal modifications are presented by the data point shapes (circle, dimethylated; triangle, acetylated). iceLogos visualize the amino acid frequency of regulated proteolytic cleavage events deduced from putative proteolysis-generated N-terminal peptides (significantly regulated, dimethylated noncanonical peptides), nonredundant N-terminal cleavage windows of the corresponding N-termini for PAN (c, d), and WT1 data sets (e, f) including the most significantly enriched gene ontology (GO) biological process (BP) terms as determined by Enrichr. The number of nonredundant, N-terminal cleavage windows used for the generation of the iceLogo is depicted on each panel, and the scissile bond is marked by a dashed line.

as marked by MaxQuant, or as invalid identification such as peptides identified with pyro-Glu modification, although they lack an N-terminal glutamic acid or glutamine residue at position 1. Thus, the final data set comprised 6170 bona fide N-terminal peptides from 1813 proteins in the PAN-treated rats and 3565 termini from 1558 proteins in the WT1 heterozygous mice (Figure S3). MANTI then further aggregated the peptides mapping to a unique position within a protein model, including the N-terminal peptides that have the same sequence but differ in their modification (e.g., acetylated and dimethylated peptides) or N-terminal peptides featuring a missed cleavage site or derived from different digestion enzymes. In both data sets, >60% of the sequences defined a unique position (Figure 3b,d). Almost 69 and 63% of the peptides in the PAN and WT1 data sets were quantified

based on the dimethyl label for quantification for both data sets. Lastly, more than 30% of the peptides in each data set feature a remapped preferred protein ID in comparison to the MaxQuant-assigned razor protein ID (which was typically the alphabetically first of multiple assigned protein IDs), allowing improved categorization (Figure 3c,e). In both data sets, 1839 N-terminal peptides from 1238 proteins (PAN) and 1605 N-terminal peptides from 1130 proteins (WT1het) matched position 1 or 2 of the preferred protein entries or within 5 amino acids of known or predicted signal peptide, transit peptide, and propeptide processing sites, which we collectively term “canonical” protein termini. As expected,⁴⁷ the majority of cytosolic protein termini starting at P1 or at P2 after the removal of the initiating Met were N-terminally acetylated, and

only a smaller proportion was modified by a dimethyl label, indicating a free amino group in vivo (Figure 3c,e).

Protein N-termini derived from known proteolytic maturation steps (signal, transit, and propeptide processing) were almost exclusively unmodified in vivo. In contrast, 3383 and 1258 N-terminal peptides matched to positions within 905 and 644 proteins in the PAN and WT1het data sets, respectively (Figures 3 and S3). The vast majority of these, in the following text termed as noncanonical termini, were mostly dimethylated and thus presumably derived from proteolytic processing, whereas a smaller fraction of N-terminal acetylated peptides may arise from unrecognized alternative translation initiation or alternative splicing.⁴⁸ Notably, we observed a higher proportion of noncanonical protein termini in the data set from PAN-treated rats (Figure 3c) compared to that from WT1 mice (Figure 3e).

According to the N-degron rule, protein termini and their modification have a strong impact on protein stability.⁴⁹ We therefore implemented an integrated auxiliary script (bufipretty.R) in MANTI to visualize the frequency of different amino acid residues as heatmaps, both collectively and separated by subcellular protein localization as annotated by the UniProt⁴⁰ database annotation (Figures S4 and S5). Acetylated N-terminal peptides were mostly found in the cytosol and nucleus-localized proteins (Figures S4a and S5a) and frequently had Met, Ala, and Ser as first amino acids, reflecting previous observations for amino acid occurrence at P1 and P2.^{50,51} In contrast, dimethylated N-termini feature a broader distribution of first amino acids (Figure S4c,d) with a bias on Ala and Ser for the noncanonical cleavages (Figures S4d and S5d), followed by Thr, Gly, and Asp. In general, dimethylated, noncanonical cleavages feature more frequently first amino acid residues classified as destabilizing.⁴⁹

Distinct Proteolytic Processes Are Associated with Chronic Kidney Injury in Both Disease Models. To identify proteolytic cleavages significantly altered in injured kidneys, MaxQuant peptide abundance ratios were parsed by MANTI and significant changes (LIMMA-moderated *t*-test $p < 0.05$ and at least 50% change in abundance) visualized by sophisticated volcano plots that indicate N-terminal peptide modification and classification (Figure 4a,b). In the more acute, chemical PAN kidney disease model, 535 (8.7%) of the N-terminal peptides showed significant changes (Figure 3a), whereas only 214 (6.0%) were found to be significantly altered in the genetic WT1 mouse model (Figure 4b). To gain insights into underlying proteolytic activities, regulated presumably protease-generated neo-N-terminal peptides (dimethylated noncanonical peptides) were aligned and the sequence surrounding the cleavage sites plotted using the iceLogo server⁵² (Figure 4c–f). Interestingly, similar iceLogos were obtained in both disease models. Depleted N-terminal peptides had an over-representation of Arg at P1 preceding the cleavage site (Figure 4c,e). The WT1 data set also indicated a higher prevalence of Trp and Asn at this position (Figure 4e), although this may in part derive from the low number of reduced cleavage events observed in heterozygous WT1 mice (27 compared to 253 in PAN-treated rats; Figure 4c). The significantly upregulated N-termini show a strong over-representation of Asn in both disease models (Figure 4d,f). In the PAN model, the cleavage logo additionally features Glu at P1 (Figure 4d). Together, this suggests that common proteolytic activities are regulated in both chronic kidney injury models. Arg-specific proteolytic cleavages appear less

prevalent, as also observed in our previous in vitro data set using PAN-treated podocyte cell cultures,⁵ whereas cleavage after Asn is induced. It is tempting to speculate that the accumulating N-terminal peptides in PAN-treated rats and WT1het might be derived from asparaginyl endopeptidases (AEPs/legumains), cysteine proteases that exhibit a clear preference for cleavage after an Asn residue at an acidic pH.⁵³ Legumain is localized in lysosomes and endocytic pathway and has previously been implicated in extracellular matrix remodeling of renal proximal tubular cells.⁵⁴ In contrast, the Arg-specific cleavage motif derived from noncanonical N-termini with reduced abundance is often (although not exclusively) observed for trypsin-like extracellular serine proteases.⁵⁵ In PAN-treated rats, Leu is additionally over-represented at the positions P2/P3 as well as Phe at P2, which is consistent with the specificity of kallikreins.⁵⁶ Interestingly, a group of gene ontology⁵⁷ (GO) terms are regulated differentially in the two animal models as evaluated by Enrichr analysis:⁵⁸ hexose biosynthetic process, glucose metabolic process, and gluconeogenesis are strongly enriched in the significantly accumulating N-termini (Figure 4f, Table S7) for WT1het mice and decreased N-termini for the PAN-treated rats (Figure 4c and Table S6). Sorbitol dehydrogenase (Sord) was one example where proteolytic N-termini were differently regulated: rat Sord (P27867) N-termini starting at position 120 (CATPPDDGNLCR) and 192 (AKAMGASQVVVIDL-SASR) were significantly depleted in the PAN treatment, whereas for the mouse Sord (Q64442), one proteolytic N-terminus starting at position 41 (SVGICGSDVHYWEHGR) significantly accumulated in the WT1het mice. Notably, in both models, the canonical N-terminus did not show a strong change in abundance, indicating that Sord protein turnover may be differentially regulated in both models. In both models, proteolysis-derived neo-N-termini of proteins involved in blood coagulation accumulated. Specifically, accumulation of neo-N-termini in the PAN model belonged to proteins associated with regulated exocytosis and protein metabolism (Table S5), whereas depleted neo-N-termini in the WT1het mice mainly comprised of fatty acid metabolism proteins (Table S8).

To identify common and distinct processing events, we first determined N-terminal peptides mapping to similar cleavage sites in homologous rat and mice proteins. The majority of the shared N-terminal peptides showed no strong changes in both data sets (Figure S6a). A subset of mostly canonical acetylated N-terminal peptides matching positions 1 and 2 was strongly depleted in both models (Figure S6b), indicating decreased protein abundance. Notably, we did not observe common induced cleavage events. A number of neo-N-termini accumulated in the PAN rats with unchanged abundance in heterozygous WT1 mice (Figure S6a,c). Among these were cleavages in cytoskeletal proteins such as actin (Actv) and serum proteins such as albumin (Alb), fibrinogen alpha (Fba)/beta (Fbg), and complement C3 (Figure S6c). Consistent with the acute, more prominent PAN-induced kidney damage, N-terminal peptides were observed, indicating the proteolytic activation of complement C3 in cisplatin-induced kidney injury in mice,⁵⁹ while the observed peptides derived from fibrinogen cleavage have been proposed as potential biomarkers for kidney disease.⁶⁰

CONCLUSIONS

Protein N-terminal peptides provide a wealth of information on co- and post-translational N-terminal modifications, protein stability, and proteolytic processing. We present MANTI as a new software tool that integrates well-established sources of protein annotation information including UniProt, LOCALIZER,⁴³ TargetP,⁴⁴ and TopFINDER⁴² with routines for statistical evaluation of quantitative changes such as limma-moderated *t*-tests³⁶ to facilitate the generation of testable hypotheses on the origin and the function of identified protein N-termini. A graphical user interface, Yoğurtlu MANTI, provides a convenient, low-threshold opportunity for N-terminome data annotation while auxiliary R scripts (bufipretty.R and polvops.R) further enable visualization and characterization of the N-termini data set in publication-quality graphs. MANTI has been designed to handle peptide identification data from the popular MaxQuant data analysis pipeline, but it is envisioned that MANTI will be expanded to accommodate identifications obtained from other common data analysis tools such as OpenMS⁶¹ and the TPP.²⁷ The source code is freely available under a Perl Artistic License v2.0 and usable for commercial and private purposes without any limitations. Finally, we demonstrate the value of MANTI for biological interpretation using two large-scale N-terminome data sets comprising >10 000 identified N-terminal peptides from two animal models of chronic kidney disease, which revealed common and distinct changes in proteolytic processing. These data sets further substantiate the importance of proteolytic processing in maintaining glomeruli function and may serve as a useful resource for further studies.⁴⁶

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acs.analchem.1c00310>.

N-terminal peptide purification workflow; screenshot of the Yoğurtlu MANTI interface to MANTI; venn diagrams with overlap of canonical and noncanonical cleavages-containing proteins; heatmaps depicting the frequency of the first amino acids for PAN N-termini; heatmaps depicting the frequency of the first amino acids for WT1 N-termini; correlation among the PAN and WT1 kidney disease models' N-termini; and main output directories and files generated by MANTI and their contents (PDF)

Full MANTI output file for PAN and WT1 datasets as well as the merged, combined N-terminal datasets and Enrichr GO BP term enrichment output (XLSX)

AUTHOR INFORMATION

Corresponding Authors

Fatih Demir – Department of Biomedicine, Aarhus University, 8000 Aarhus C, Denmark; Central Institute for Engineering, Electronics and Analytics, ZEA-3, Forschungszentrum Jülich GmbH, 52425 Jülich, Germany; orcid.org/0000-0002-5744-0205; Phone: +45 2066 5308; Email: fatih.demir@biomed.au.dk

Pitter F. Huesgen – Central Institute for Engineering, Electronics and Analytics, ZEA-3, Forschungszentrum Jülich GmbH, 52425 Jülich, Germany; Cologne Excellence Cluster Cellular Stress Response in Aging-Associated Diseases (CECAD), Medical Faculty and University Hospital,

Institute of Biochemistry, Department of Chemistry, University of Cologne, 50931 Cologne, Germany; orcid.org/0000-0002-0335-2242; Phone: +49 246 61 6873; Email: p.huesgen@fz-juelich.de

Authors

Jayachandran N. Kizhakkedathu – Centre for Blood Research, Department of Pathology & Laboratory Medicine, School of Biomedical Engineering, Department of Chemistry, University of British Columbia, Vancouver V6T 1Z3, British Columbia, Canada; orcid.org/0000-0001-7688-7574

Markus M. Rinschen – Department of Biomedicine, Aarhus University, 8000 Aarhus C, Denmark; III. Department of Medicine, University Medical Center Hamburg-Eppendorf, 20251 Hamburg, Germany

Complete contact information is available at:

<https://pubs.acs.org/doi/10.1021/acs.analchem.1c00310>

Author Contributions

F.D. and P.F.H. conceived the project. F.D. wrote the software. M.M.R. and P.F.H. designed the experiments. F.D. analyzed the data. J.N.K. provided novel reagents. The manuscript was written by F.D. and P.F.H. and edited by all authors. All authors have approved the final version of the manuscript.

Notes

The authors declare no competing financial interest. MANTI is available for free (Perl Artistic License v2.0) via the project home page: <https://MANTI.sourceforge.io>. A 61 page documentation, as well as a short tutorial, is available through the accompanying PDF manual within each MANTI release.

ACKNOWLEDGMENTS

P.F.H. acknowledges support by the European Research Council under the European Union's Horizon 2020 research and innovation program (starting grant ProPlantStress, ID 639905) and the Deutsche Forschungsgemeinschaft (DFG, FOR 2743 mechanical stress protection, HU 1756/3-1). M.M.R. acknowledges support by the Young Investigator Award from the Novo Nordisk Foundation (grant number NNF19OC0056043), Aarhus Universitet Forskningsfond and the DFG (FOR 2743 mechanical stress protection, RI 2811/2). J.N.K. acknowledges the funding from Canadian Institutes of Health Research. The funding bodies had no role in the study design, collection, analysis, and interpretation of data, writing of the manuscript, and in the decision to submit the manuscript for publication. We would like to thank Nadine Wettengl (Juelich) for technical support in the MS sample preparation. The WT1 heterozygous knockout mice were a kind gift of Jordan Kreidberg (Boston). We thank Florian Grahmmer (Hamburg-Eppendorf) for providing the glomeruli from PAN-treated rats and Martin Kann (Cologne) for providing kidney samples from WT1 heterozygous mice.

REFERENCES

- (1) Perrar, A.; Dissmeyer, N.; Huesgen, P. F. *J. Exp. Bot.* **2019**, *70*, 2021–2038.
- (2) Klein, T.; Eckhard, U.; Dufour, A.; Solis, N.; Overall, C. M. *Chem. Rev.* **2018**, *118*, 1137–1168.
- (3) Saido, T.; Leissring, M. A. *Cold Spring Harbor Perspect. Med.* **2012**, *2*, No. a006379.
- (4) Kessenbrock, K.; Plaks, V.; Werb, Z. *Cell* **2010**, *141*, 52–67.
- (5) Rinschen, M. M.; Hoppe, A.; Grahmmer, F.; Kann, M.; Völker, L. A.; Schurek, E.; Binz, J.; Höhne, M.; Demir, F.; Malisic, M.; Huber,

- T. B.; Kurschat, C.; Kizhakkeedathu, J. N.; Schermer, B.; Huesgen, P. F.; Benzing, T. J. *Am. Soc. Nephrol.* **2017**, *28*, 2867–2878.
- (6) Dufour, A.; Overall, C. M. *Trends Pharmacol. Sci.* **2013**, *34*, 233–242.
- (7) Cox, J.; Mann, M. *Nat. Biotechnol.* **2008**, *26*, 1367–1372.
- (8) Ludwig, C.; Gillet, L.; Rosenberger, G.; Amon, S.; Collins, B. C.; Aebersold, R. *Mol. Syst. Biol.* **2018**, *14*, 1–23.
- (9) Demir, F.; Niedermaier, S.; Kizhakkeedathu, J. N.; Huesgen, P. F. TAILS Profiling of Protein N Termini and Their Modifications in Complex Samples. In *Methods in Molecular Biology*; Schilling, O., Ed.; Springer, 2017.
- (10) Niedermaier, S.; Huesgen, P. F. *Biochim. Biophys. Acta, Proteins Proteomics* **2019**, 1867, No. 140138.
- (11) Bogaert, A.; Gevaert, K. *Expert Rev. Proteomics* **2020**, *17*, 581–594.
- (12) Kleifeld, O.; Doucet, A.; Auf Dem Keller, U.; Prudova, A.; Schilling, O.; Kainthan, R. K.; Starr, A. E.; Foster, L. J.; Kizhakkeedathu, J. N.; Overall, C. M. *Nat. Biotechnol.* **2010**, *28*, 281–288.
- (13) Venne, A. S.; Vögtle, F. N.; Meisinger, C.; Sickmann, A.; Zahedi, R. P. J. *Proteome Res.* **2013**, *12*, 3823–3830.
- (14) Gevaert, K.; Goethals, M.; Martens, L.; Van Damme, J.; Staes, A.; Thomas, G. R.; Vandekerckhove, J. *Nat. Biotechnol.* **2003**, *21*, 566–569.
- (15) Chang, C.-H.; Chang, H.-Y.; Rappsilber, J.; Ishihama, Y. *Mol. Cell. Proteomics* **2021**, No. 100003.
- (16) Weng, S. S. H.; Demir, F.; Ergin, E. K.; Dirnberger, S.; Uozio, A.; et al. *Mol. Cell. Proteomics* **2019**, *18*, 2335–2347.
- (17) Demir, F.; Niedermaier, S.; Villamor, J. G.; Huesgen, P. F. *New Phytol.* **2018**, 936–943.
- (18) Du, X.; Lu, J.; Yang, L.; Bao, H.; Zhang, L.; Yan, G.; Fang, C.; Lu, H. *Anal. Chem.* **2020**, *92*, 8315–8322.
- (19) Sun, M.; Liang, Y.; Li, Y.; Yang, K.; Zhao, B.; Yuan, H.; Li, X.; Zhang, X.; Liang, Z.; Shan, Y.; Zhang, L.; Zhang, Y. *Anal. Chem.* **2020**, *92*, 567–572.
- (20) Soh, W. T.; Demir, F.; Dall, E.; Perrar, A.; Dahms, S. O.; Kuppusamy, M.; Brandstetter, H.; Huesgen, P. F. *Anal. Chem.* **2020**, 2961–2971.
- (21) Omasits, U.; Ahrens, C. H.; Müller, S.; Wollscheid, B. *Bioinformatics* **2014**, *30*, 884–886.
- (22) Bateman, A.; Martin, M. J.; O'Donovan, C.; Magrane, M.; Alpi, E.; Antunes, R.; Bely, B.; Bingley, M.; Bonilla, C.; Britto, R.; Bursteinas, B.; Bye-Ajee, H.; Cowley, A.; Da Silva, A.; De Giorgi, M.; Dogan, T.; Fazzini, F.; Castro, L. G.; Figueira, L.; Garmiri, P.; Georgiadiou, G.; Zhang, J.; et al. *Nucleic Acids Res.* **2017**, *45*, D158–D169.
- (23) Ivankov, D. N.; Bogatyreva, N. S.; Hönigschmid, P.; Dislich, B.; Hög, S.; Kuhn, P.-H.; Frishman, D.; Lichtenthaler, S. F. *Nucleic Acids Res.* **2013**, *41*, W459–W464.
- (24) Nilse, L.; Avcı, D.; Heisterkamp, P.; Serang, O.; Lemberg, M. K.; Schilling, O. *Biochim. Biophys. Acta, Proteins Proteomics* **2016**, 1864, 1363–1371.
- (25) Auf Dem Keller, U.; Overall, C. M. *Biol. Chem.* **2012**, *393*, 1477–1483.
- (26) Nesvizhskii, A. I.; Aebersold, R. *Mol. Cell. Proteomics* **2005**, *4*, 1419–1440.
- (27) Deutsch, E. W.; Mendoza, L.; Shteynberg, D.; Slagel, J.; Sun, Z.; Moritz, R. L. *Proteomics: Clin. Appl.* **2015**, *9*, 745–754.
- (28) Tyanova, S.; Temu, T.; Cox, J. *Nat. Protoc.* **2016**, *11*, 2301–2319.
- (29) Hofsetz, E.; Demir, F.; Szczepanowska, K.; Kukat, A.; Kizhakkeedathu, J. N.; Trifunovic, A.; Huesgen, P. F. *Mol. Cell. Proteomics* **2020**, *19*, 1330–1345.
- (30) Mielke, K.; Wagner, R.; Mishra, L. S.; Demir, F.; Perrar, A.; Huesgen, P. F.; Funk, C. J. *Exp. Bot.* **2020**, No. eraa550.
- (31) Dall, E.; Zauner, F. B.; Soh, W. T.; Demir, F.; Dahms, S. O.; Cabrele, C.; Huesgen, P. F.; Brandstetter, H. *J. Biol. Chem.* **2020**, 13047–13064.
- (32) Shankland, S. J. *Kidney Int.* **2006**, *69*, 2131–2147.
- (33) Vega-Warner, V.; Ransom, R. F.; Vincent, A. M.; Brosius, F. C.; Smoyer, W. E. *Kidney Int.* **2004**, *66*, 1881–1889.
- (34) Schumacher, V. A.; Schlötzer-Schrehardt, U.; Karumanchi, S. A.; Shi, X.; Zaia, J.; Jeruschke, S.; Zhang, D.; Pavenstaedt, H.; Drenckhan, A.; Amann, K.; Ng, C.; Hartwig, S.; Ng, K. H.; Ho, J.; Kreidberg, J. A.; Taglienti, M.; Royer-Pokora, B.; Ai, X. J. *Am. Soc. Nephrol.* **2011**, *22*, 1286–1296.
- (35) Rappsilber, J.; Mann, M.; Ishihama, Y. *Nat. Protoc.* **2007**, *2*, 1896–1906.
- (36) Ritchie, M. E.; Phipson, B.; Wu, D.; Hu, Y.; Law, C. W.; Shi, W.; Smyth, G. K. *Nucleic Acids Res.* **2015**, *43*, No. e47.
- (37) R Core Team. R: A Language and Environment for Statistical Computing; R Core Team: Vienna, Austria, 2017.
- (38) Wickham, H. *Ggplot2: Elegant Graphics for Data Analysis*; Springer-Verlag: New York, 2009.
- (39) Lex, A.; Gehlenborg, N.; Strobelt, H.; Vuilleumot, R.; Pfister, H. *IEEE Trans. Visualization Comput. Graphics* **2014**, *20*, 1983–1992.
- (40) Perez-Riverol, Y.; Csordas, A.; Bai, J.; Bernal-Llinares, M.; Hewapathirana, S.; Kundu, D. J.; Inuganti, A.; Griss, J.; Mayer, G.; Eisenacher, M.; Pérez, E.; Uszkoreit, J.; Pfeuffer, J.; Sachsenberg, T.; Yilmaz, S.; Tiwary, S.; Cox, J.; Audain, E.; Walzer, M.; Jarnuczak, A. F.; Ternent, T.; Brazma, A.; Vizcaíno, J. A. *Nucleic Acids Res.* **2019**, *47*, D442–D450.
- (41) Wall, L.; Christiansen, T.; Orwant, J. *Programming Perl*, 3rd ed.; O'Reilly Media, Inc., 2000.
- (42) Fortelny, N.; Yang, S.; Pavlidis, P.; Lange, P. F.; Overall, C. M. *Nucleic Acids Res.* **2015**, *43*, D290–D297.
- (43) Sperschneider, J.; Catanzariti, A. M.; Deboer, K.; Petre, B.; Gardiner, D. M.; Singh, K. B.; Dodds, P. N.; Taylor, J. M. *Sci. Rep.* **2017**, *7*, No. 44598.
- (44) Armenteros, J. J. A.; Salvatore, M.; Emanuelsson, O.; Winther, O.; Von Heijne, G.; Eloffsson, A.; Nielsen, H. *Life Sci. Alliance* **2019**, *2*, No. e201900429.
- (45) Nesvizhskii, A. I.; Keller, A.; Kolker, E.; Aebersold, R. *Anal. Chem.* **2003**, *75*, 4646–4658.
- (46) Rinschen, M. M.; Huesgen, P. F.; Koch, R. E. *Am. J. Physiol.: Renal Physiol.* **2018**, *315*, F1812–F1816.
- (47) Dinh, T. V.; Bienvenut, W. V.; Linster, E.; Feldman-Salit, A.; Jung, V. A.; Meinel, T.; Hell, R.; Giglione, C.; Wirtz, M. *Proteomics* **2015**, *15*, 2426–2435.
- (48) Yeom, J.; Ju, S.; Choi, Y.; Paek, E.; Lee, C. *Sci. Rep.* **2017**, *7*, No. 6599.
- (49) Varshavsky, A. *Proc. Natl. Acad. Sci. U.S.A.* **2019**, *116*, 358–366.
- (50) auf dem Keller, U.; Prudova, A.; Eckhard, U.; Fingleton, B.; Overall, C. M. *Sci. Signaling* **2013**, *6*, No. rs2.
- (51) Lange, P. F.; Huesgen, P. F.; Nguyen, K.; Overall, C. M. *J. Proteome Res.* **2014**, *13*, 2028–2044.
- (52) Colaert, N.; Helsens, K.; Martens, L.; Vandekerckhove, J.; Gevaert, K. *Nat. Methods* **2009**, *6*, 786–787.
- (53) Dall, E.; Brandstetter, H. *Biochimie* **2016**, *122*, 126–150.
- (54) Morita, Y.; Araki, H.; Sugimoto, T.; Takeuchi, K.; Yamane, T.; Maeda, T.; Yamamoto, Y.; Nishi, K.; Asano, M.; Shirahama-Noda, K.; Nishimura, M.; Uzu, T.; Hara-Nishimura, I.; Koya, D.; Kashiwagi, A.; Ohkubo, I. *FEBS Lett.* **2007**, *581*, 1417–1424.
- (55) Gosalia, D. N.; Salisbury, C. M.; Ellman, J. A.; Diamond, S. L. *Mol. Cell. Proteomics* **2005**, *4*, 626–636.
- (56) Li, H.-X.; Hwang, B.-Y.; Laxmikanthan, G.; Blaber, S. I.; Blaber, M.; Golubkov, P. A.; Ren, P.; Iverson, B. L.; Georgiou, G. *Protein Sci.* **2008**, *17*, 664–672.
- (57) The Gene Ontology Consortium; Carbon, S.; Douglass, E.; Good, B. M.; Unni, D. R.; Harris, N. L.; Mungall, C. J.; Basu, S.; Chisholm, R. L.; Dodson, R. J.; Hartline, E.; Fey, P.; Thomas, P. D.; Albou, L.-P.; Ebert, D.; Kesling, M. J.; Mi, H.; Muruganujan, A.; Huang, X.; Mushayahama, T.; LaBonte, S. A.; Elser, J.; et al. *Nucleic Acids Res.* **2021**, *49*, D325–D334.
- (58) Kuleshov, M. V.; Jones, M. R.; Rouillard, A. D.; Fernandez, N. F.; Duan, Q.; Wang, Z.; Koplev, S.; Jenkins, S. L.; Jagodnik, K. M.; Lachmann, A.; McDermott, M. G.; Monteiro, C. D.; Gundersen, G. W.; Ma'ayan, A. *Nucleic Acids Res.* **2016**, *44*, W90–W97.

(59) Späth, M. R.; Bartram, M. P.; Palacio-Escat, N.; Hoyer, K. J. R.; Debes, C.; Demir, F.; Schroeter, C. B.; Mandel, A. M.; Grundmann, F.; Ciarimboli, G.; Beyer, A.; Kizhakkedathu, J. N.; Brodesser, S.; Göbel, H.; Becker, J. U.; Benzing, T.; Schermer, B.; Höhne, M.; Burst, V.; Saez-Rodriguez, J.; Huesgen, P. F.; Müller, R.-U.; Rinschen, M. M. *Kidney Int.* **2018**, 1–17.

(60) Wang, H.; Zheng, C.; Lu, Y.; Jiang, Q.; Yin, R.; Zhu, P.; Zhou, M.; Liu, Z. *Clin. J. Am. Soc. Nephrol.* **2017**, 12, 1922–1929.

(61) Röst, H. L.; Sachsenberg, T.; Aiche, S.; Bielow, C.; Weisser, H.; Aicheler, F.; Andreotti, S.; Ehrlich, H.-C.; Gutenbrunner, P.; Kenar, E.; Liang, X.; Nahnsen, S.; Nilse, L.; Pfeuffer, J.; Rosenberger, G.; Rurik, M.; Schmitt, U.; Veit, J.; Walzer, M.; Wojnar, D.; Wolski, W. E.; Schilling, O.; Choudhary, J. S.; Malmström, L.; Aebersold, R.; Reinert, K.; Kohlbacher, O. *Nat. Methods* **2016**, 13, 741–748.