Journal of proteome research

Article

pubs.acs.org/jpr

# The Human Dental Pulp Proteome and N-Terminome: Levering the Unexplored Potential of Semitryptic Peptides Enriched by TAILS to Identify Missing Proteins in the Human Proteome Project in Underexplored Tissues

Ulrich Eckhard,[†,‡] Giada Marino,[†,‡] Simon R. Abbey,[†,‡] Grace Tharmarajah,[§] Ian Matthew,[‡] and Christopher M. Overall*[,†,‡,∥]

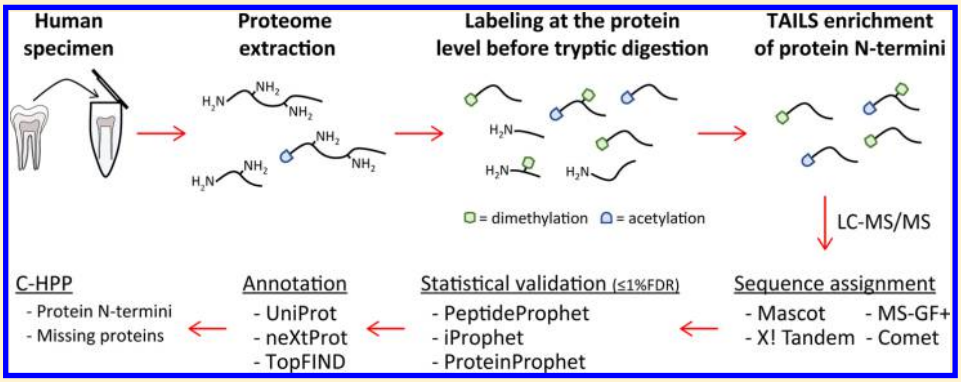
[†]Centre for Blood Research, University of British Columbia, Vancouver, British Columbia V6T 1Z3, Canada

[‡]Department of Oral Biological and Medical Sciences, Faculty of Dentistry, University of British Columbia, Vancouver, British Columbia V6T 1Z3, Canada

[§]Department of Medical Genetics, Faculty of Medicine, University of British Columbia, Vancouver, British Columbia V6T 1Z4, Canada

[∥]Department of Biochemistry and Molecular Biology, University of British Columbia, Vancouver, British Columbia V6T 1Z4, Canada

Supporting Information




**ABSTRACT:** An underexplored yet widespread feature of the human proteome is the proteolytic proteoforms of proteins. We used terminal amine isotopic labeling of substrates (TAILS), a high-content N-terminal positional proteomics technique, for in-depth characterization of the human dental pulp proteome from its N-terminome and to provide data for the Chromosome-centric Human Proteome Project (C-HPP). Dental pulp is a unique connective tissue maintaining tooth sensation and structure by supporting a single cell layer of odontoblasts that synthesize mineralization-competent dentine extracellular matrix. Therefore, we posited pulp to be a rich source of unique tissue-specific proteins and hence an abundant source of "missing" proteins as defined by neXtProt. From the identified 4332 proteins (false discovery rate (FDR) ≤ 0.7%), 21 528 unique peptides (FDR ≤ 1.0%) and 9079 unique N-termini, we analyzed N-terminal methionine excision, co- and posttranslational N$\alpha$-acetylation, protein maturation, and proteolytic processing. Apart from 227 candidate alternative translation initiation sites, most identified N-termini (78%) represented proteolytic processing and mechanism-informative internal neo-N-termini, confirming a pervasive amount of proteolytic-processing generated proteoforms in vivo. Furthermore, we identified 17 missing protein candidates for the C-HPP, highlighting the importance of using (i) less studied human specimens and (ii) orthogonal proteomic approaches such as TAILS to map the human proteome. The mass spectrometry raw data and metadata have been deposited to ProteomeXchange with the PXD identifier <PXD002264>.

**KEYWORDS:** *dental pulp, dentin, odontoblasts, teeth, TAILS, N-terminome, N-termini, missing proteins, positional proteomics*

## INTRODUCTION

An integral central plank of the HUPO Human Proteome Project is the Chromosome-centric Human Proteome Project (C-HPP; www.c-hpp.org) that aims to experimentally identify and characterize at least one protein product for each of the ~20 100 human protein-coding genes.[1] Currently, more than 2500 proteins are still awaiting validation on the protein level by

**Special Issue:** The Chromosome-Centric Human Proteome Project 2015

**Received:** June 21, 2015
**Published:** August 10, 2015




DOI: 10.1021/acs.jproteome.5b00579
J. Proteome Res. 2015, 14, 3568−3582

either mass spectrometry or antibodies, and are thus classified as "missing" (www.nextprot.org).[2] By definition, missing proteins only have evidence on a transcriptomic level (PE2), or are inferred from homology (PE3) or are predicted by sequence (PE4). Further, for another 600 entries it is even unclear if they are actually protein-coding genes, and consequently they are assigned as "uncertain" (PE5).[3,4] The high number of missing proteins in the human proteome (>12%) may be surprising considering the many large scale proteomics studies published, but can be rationalized as many proteins are (i) only expressed in either inaccessible or relatively unexplored organs, tissues, or cell types, which are therefore often jointly referred to as unusual or rare human specimens; (ii) only translated during distinct developmental stages; (iii) activated only under distinct stress conditions; or are (iv) present, but below current detection limits within complex tissues and biofluids. Furthermore, (v) especially membrane-embedded proteins often possess unsuitable physicochemical properties and lack tryptic cleavage sites rendering their handling unwieldy and their detection unlikely. Lastly, (vi) many protein (sub)families share high sequence homology and thus identical peptide-spectrum matches (PSMs), leading to their exclusion during analysis as by the law of parsimony only one representative is chosen for the final nonredundant protein list.[5] Consequently, even by using state-of-the-art technologies and methodologies, detection and quantification capacities are still restricted on a large-scale proteomics level, and complementary strategies with orthogonal strengths are needed to fill this cavity.

In our recent analyses of rarely investigated cells and tissues to seek missing proteins for the HPP, we have uncovered with high confidence six proteins in erythrocytes[6] and ten proteins in platelets[7] by using a combined two-pronged strategy: (i) exploiting unusual tissues or specialized cell types in order to identify the proteins that contribute to their uniqueness in terms of anatomy, location and function; and (ii) identifying missing proteins using the underexploited power of searching for semitryptic peptides that are commonly overlooked in typical database searches. In particular, semitryptic peptides originate from protein natural N-termini, so also providing valuable protein-characterization informative data, but mainly from surprisingly greater numbers of neo-N and neo-C termini generated from internal proteolytically processed sites. We reasoned that many such neo-termini will have improved properties for peptide ionization and fragmentation due to differences in mass, charge, and hence $m/z$, and amino acid distribution, when compared to their spanning tryptic peptides (Figure 1). Thereby, semitryptic peptides can be used to identify certain segments in a protein that otherwise may be recalcitrant to conventional shotgun proteomics. Our evidence suggests these segments, on occasion, may actually form the only part of proteins amenable to proteomics. However, a key reason that semitryptic peptides are commonly overlooked is that they are present in relatively low abundance in shotgun proteomics analyses. Therefore, as for successful analysis of most post-translational modifications, enrichment techniques are required to improve breadth and depth of their coverage. One such technique is the highly successful, high content approach, terminal amine isotopic labeling of substrates (TAILS).[8]

Protein N-termini typically originate directly from translation or proteolytic maturation, e.g., N-terminal methionine excision, signal or transit peptide cleavage, prodomain removal, and internal maturation cleavage such as after disulfide bond formation to generate two linked protein chains with four

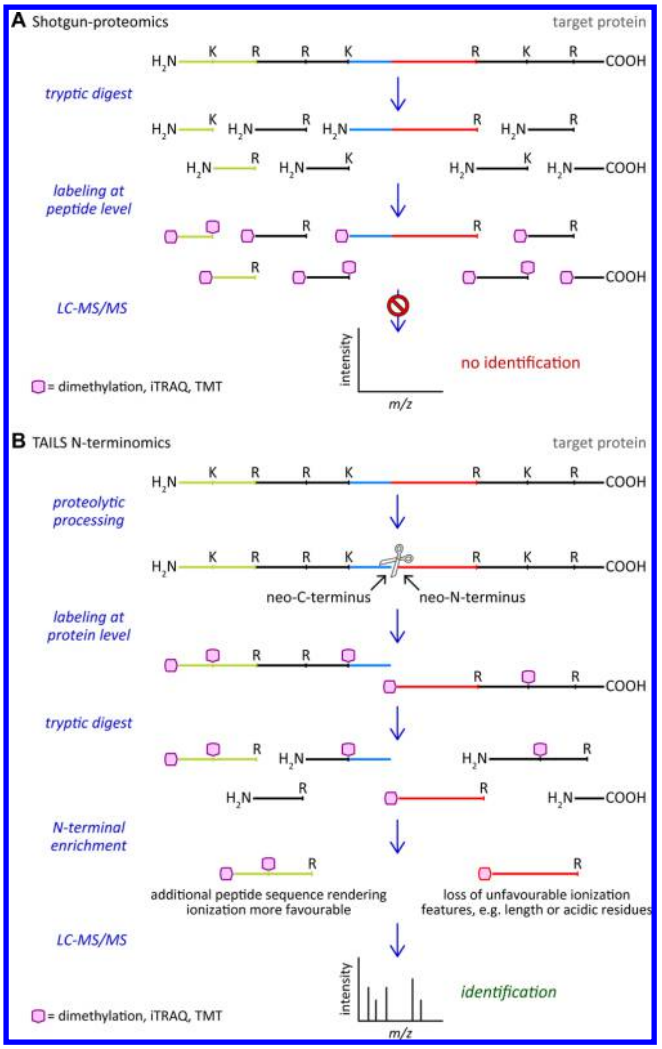



**Figure 1.** Schematic representation highlighting the application of TAILS N-terminomics in identifying proteins where fully tryptic peptides are recalcitrant to LC−MS/MS, especially low abundance proteins or those with unfavorable ionization or fragmentation properties that are typically missed by conventional shotgun proteomics approaches. (A) Problematic peptides generated by trypsin digestion in conventional shotgun proteomics workflows. (B) TAILS workflow showing the corresponding semitryptic counterparts that have altered length and amino acid distribution and thus higher probability for improved ionization and fragmentation, and thus protein identification. In TAILS, trypsin cuts with ArgC specificity due to the blocked lysine residues.

termini. The other large category of protein N-termini affecting almost all proteins results from specific proteolytic processing within the mature protein chain, frequently altering the function of bioactive molecules. Note, processing is distinct from general degradative cleavage events resulting in protein breakdown. Indeed 44% of normal murine skin proteins start distal to the expected protein start sites,[9] and 68% and 77% of all identified proteins in human erythrocytes[6] and human platelets,[7] respectively, possess stable neo-N-termini. Thus, an underexplored aspect of the human proteome is the prevalence, characterization and functional outcome of such proteolytic proteoforms generated by precision proteolysis. TAILS is the most amenable N-terminomics approach for simultaneous high-throughput identification of natural and neo-N-termini, requiring few analyses per sample, and with proven reliability


DOI: 10.1021/acs.jproteome.5b00579
J. Proteome Res. 2015, 14, 3568−3582

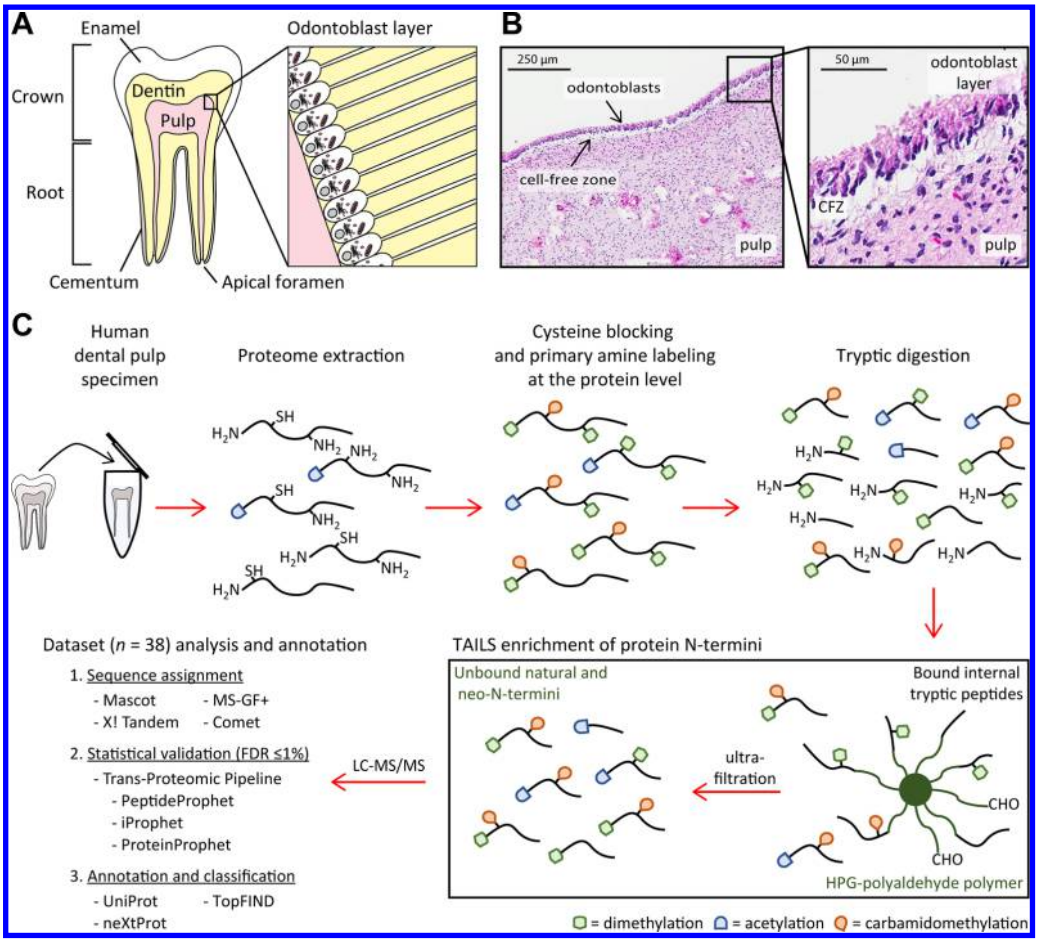



**Figure 2.** (A) Schematic representation of human tooth structure. The anatomical parts of a tooth include from top to bottom: enamel, dentin, pulp, and cementum. Whereas the bone-like dentin tissue constitutes the largest portion of a tooth, the central soft tissue known as the dental pulp, contains all the nerves and blood vessels fundamental for keeping a tooth healthy, and which hyphenate with the exterior periodontium via the apical foramen. The heavily calcified enamel and bone-like cementum protect the tooth crown above and the root below the gingival margin, respectively. With high secretory capacity of dentin proteins, large specialized columnar cells known as odontoblasts are located at the pulp dentin interface. Odontoblasts display long cellular processes that pass through the entire dentin depth to the cemental or enamel interfaces. (B) Haematoxylin and eosin staining of histological section from human dental pulp. Three zones are present: the odontoblast layer in the outermost region, the cell free zone (CFZ), and the cell-rich zone that merges into the pulp core. (C) Pulp proteomics and N-terminomics workflow. Dental pulps were obtained immediately from fresh surgical extractions of wisdom teeth from nine patients. Pulp proteomes were harvested and processed following the standard TAILS N-terminomics protocol. 38 LC−MS/MS data sets were collected and all were analyzed using four database search engines and the Trans-Proteomic Pipeline for statistical analysis for peptide identification FDR ≤ 1.0%. Annotation was performed using UniProt, neXtProt, and the knowledgebase TopFIND v3.0.

and robustness shown from multiple laboratories in many different applications.[10−12] Primary amine labeling is performed at the whole protein level, overhauling trypsin to cut with ArgC specificity and thus increasing average peptide lengths of otherwise shorter semitryptic peptides. Whereas this can lead to compressed spectra from longer peptides that can result sometimes in decreased probabilities of identification, many other peptides remain identifiable with high confidence so more than counterbalancing these opposite considerations. Thus, due to the inherently different physicochemical properties of semispecific N-terminal peptides compared to fully tryptic internal peptides typically targeted by shotgun proteomics, TAILS N-terminomics represents an appealing approach for the identification of proteins so far missed by conventional shot-gun approaches.

One unusual tissue that heretofore has largely escaped proteomic analysis is the dental pulp, the soft connective tissue component of the pulpo-dentin complex confined within the dental pulp chamber and root canals and colloquially known as the "nerve" of the tooth (Figure 2A). A single layer of specialized odontoblasts (Figure 2B) line the dentin atop the pulp, and extend into the dentin as long thin processes within the dentinal tubules. Besides fibroblasts, which constitute the predominant cell type and are responsible for the synthesis of pulpal structural extracellular matrix, the pulp contains endothelial cells, and immunocompetent cells such as macrophages, dendritic cells and polymorphonuclear leukocytes.[13−15] Furthermore, the presence of undifferentiated mesenchymal cells in pulp provides regenerative potential, synthesizing reparative dentin when dental caries or trauma stimulate, damage or infect vital pulp tissue.[16,17] The primary function of pulp is the deposition of the dentin layer during tooth development, but despite dentin mineralization, remains throughout life innervated and vascularized. Dental pulp is the only vital tissue remaining in the highly mineralized tooth, and hence provides sensation to teeth and can repair dentin from within. On tooth eruption, new functions segue allowing pulp to mediate sensitivity and to detect thermal, pH, osmolarity, chemical and pressure changes or damage. However, the sensory apparatus for this is neither understood


DOI: 10.1021/acs.jproteome.5b00579
J. Proteome Res. 2015, 14, 3568−3582

nor are the receptors precisely known, but is thought to involve the odontoblast processes at exposed dentine.

Unsurprisingly, there are few large-scale proteomics analyses of pulp tissue to date,[18,19] and only two studies annotated the human pulp proteome, with just 96[20] and 342[21] proteins identified by 2D-PAGE-MS-MS and 2D-LC−MS/MS, respectively. Here, we report the targeted analysis of the human dental pulp proteome using TAILS N-terminomics combined with a pre-TAILS shotgun-like analysis, which allowed us to capture both naturally blocked and unblocked protein N-termini, and to monitor pervasive proteolytic processing in a healthy human tissue. By analyzing this virtually proteomically unstudied human tissue we: (i) identified 174 previously unidentified proteins (ProteinProphet probability of 0.95) of which 17 candidates fulfilled the highly stringent criteria as required by C-HPP, and (ii) established a general workflow suitable for the in-depth and high-throughput analysis of the human proteome and N-terminome in human specimens as envisioned by the HPP.

## EXPERIMENTAL PROCEDURES

### Dental Pulp Collection and Proteome Preparation

Healthy dental pulps were collected from nine patients, five women and four men, within 5−10 min of routine prophylactic extraction of healthy wisdom teeth without dental caries. Before surgery, written informed consent was obtained from the patient according to a protocol approved by the University of British Columbia Clinical Research Ethics Board (UBC CREB). Extracted teeth were partially sectioned vertically with a high-speed dental turbine and under water spray to avoid heating or traumatizing the pulp tissue. The teeth were then split into two pieces with a dental elevator exposing the pulp, which was removed using sterile curettes and barbed broaches. The pulp was immediately transferred into 250 $\mu$L 8 M guanidine hydrochloride, frozen on dry ice, and stored at −80 °C for a maximum period of 1 month. Care was taken to prevent adventitious proteolysis occurring by incorporation of protease inhibitor cocktails and tissue handling at 0−4 °C whenever feasible.[22] Dental pulp specimens from each of the nine patients were separately homogenized on ice using an Ultra-Turrax tissue homogenizer (IKA Works, Inc.). Three rounds of protein extraction were performed, supernatants were pooled, and proteins precipitated by chloroform/methanol.[23] Pellets were redissolved in 0.5 mL of 8 M guanidine hydrochloride and protein concentrations were determined using 1:10 dilutions (in $ddH_2O$) and Bradford assay (Bio-Rad) with bovine serum albumin as a standard.

### Histological Analysis

For histological analysis, one fresh dental pulp was immediately transferred into 4% paraformaldehyde in phosphate buffered saline and fixed overnight at 4 °C. The sample was subsequently washed four times with phosphate buffered saline, dehydrated through graded series of ethanol and cleared with xylene. The pulp sample was then embedded in paraffin blocks and cut into serial sections of 12-$\mu$m thickness. Haematoxylin and eosin (HE) staining and digital image recording were performed by Wax-it Histology Services at the University of British Columbia.

### TAILS N-Terminomics

Protein N-termini enrichment by TAILS was performed as described previously.[22] Aliquots of approximately 1.0 mg of nonfractionated dental pulp proteome were diluted to 4 M guanidine hydrochloride and reduced in 5 mM dithiothreitol (30 min, 65 °C) prior to cysteine carbamidomethylation using 10 mM iodoacetamide (45 min, room temperature in the dark). Excess blocking reagent was quenched by adding 10 mM dithiothreitol (30 min, room temperature). The pH was subsequently adjusted to 6.5 for reductive dimethylation of primary amines with 40 mM isotopically heavy formaldehyde ($^{13}CD2$ in $D_2O$; Cambridge Isotopes) and 20 mM sodium cyanoborohydride (overnight, 37 °C); contrary to conventional shotgun proteomics, blocking of primary amines is performed in TAILS at the whole protein level and thus before tryptic digestion. In this way all free protein N-termini are labeled enabling easy discrimination from any tryptic peptide carry over in the analyses. After overnight incubation, additional 20 mM heavy formaldehyde and 20 mM cyanoborohydride were added (2 h, 37 °C) to ensure completion of amine labeling. Reactions were subsequently quenched using 100 mM Tris-HCl, pH 6.8 (30 min, 37 °C) and cleaned-up by chloroform/methanol precipitation.[23] Protein pellets were resolubilized in a small volume of 50 mM NaOH and pH-neutralized using 100 mM 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid, pH 7.5 to 250 $\mu$L, diluted 1:1 with HPLC-grade water, and digested with mass spectrometry-grade trypsin (Trypsin Gold, Promega) at a proteome:enzyme ratio of 100:1 (w/w). After overnight incubation at 37 °C, digestion efficiency was checked by SDS-PAGE, and if necessary, more trypsin was added (2 h, 37 °C). An aliquot of 50 $\mu$g of tryptic digest was saved for shotgun-like analysis (preTAILS) before adjusting the samples to pH 6.5 and coupling trypsin-generated internal and C-terminal peptides to a 5-fold excess (w/w) of water-soluble HPG-ALD polymer (http://flintbox.com/public/project/1948/) by incubation with 20 mM sodium cyanoborohydride (37 °C, overnight), pH 6.8. Reactions were quenched by adding 100 mM tris-(hydroxymethyl)aminomethane buffer (pH 6.8, 30 min, 37 °C), and unbound peptides representing naturally blocked or experimentally labeled N-terminal peptides were recovered by ultrafiltration. Samples prepared in this manner are designated TAILS samples whereas aliquots of labeled peptides taken before internal tryptic peptide removal are designated preTAILS. Both preTAILS and TAILS samples were subsequently desalted using C18 StageTips,[24] flash frozen in liquid nitrogen, and stored at −80 °C until LC−MS/MS analysis.

### Mass Spectrometry

Purified peptide samples were analyzed using a quadrupole time-of-flight mass spectrometer (Accurate Mass G6550A Q-TOF; Agilent), coupled online to an Agilent 1200 Series nanoflow HPLC using a Chip Cube nanospray ionization interface (Agilent). A high capacity HPLC-Chip with 160 nL enrichment column and a 0.075 mm × 150 mm analytical column containing Zorbax 300SB-C18 5 $\mu$m stationary phase (Agilent) was used, and the thermostat temperature was set at 6 °C. Each sample was automatically loaded on the enrichment column at flow rate 4 $\mu$L/min of buffer A (0.1% formic acid) and at 4 $\mu$L injection flush volume. After that, a 110.2 min gradient was established with the nanopump at 300 nL/min from 0% to 5% buffer B (99.9% acetonitrile, 0.1% formic acid) over 2 min, then from 5% to 45% buffer B in the next 78 min, then increased to 60% over 10 min period, further increased to 95% buffer B over 0.1 min, held at 95% for 20 min, and then reduced to 3% buffer B for 0.1 min to recondition the column for the next analysis. Peptides were ionized by ESI (1.8 kV), and mass spectrometry analysis was performed in positive polarity with precursor ions detected from 300 to 2000 $m/z$. The top three ions per scan were selected for



DOI: 10.1021/acs.jproteome.5b00579
J. Proteome Res. 2015, 14, 3568−3582

CID using a narrow exclusion window of 1.3 amu and at a MS/MS scan rate of two spectra per second. Collision energy was calculated automatically depending on the charge state of the parent ions, and precursor ions were then excluded from further CID for 30 s. The entire LC−MS system was run by Mass Hunter version B.02.01 (Agilent).

### Data Analysis

Acquired MS/MS raw data were converted to mgf and mzXML files using MSConvert,[25] and spectra were matched to peptide sequences in the human UniProt protein database (release 2013_10) using four different search engines, namely Mascot v2.4,[26] X! TANDEM CYCLONE TPP 2011.12.01.1,[27] MS-GF+ v10072,[28] and Comet 2015.01 rev 0.[29] The expected cleavage pattern was set to semi-ArgC and allowed for two missed cleavages. Search parameters included 20 ppm tolerance for MS1 and 0.25 Da or 50 ppm for MS2, fixed lysine dimethylation (+34.0631 Da), fixed cysteine carbamidomethylation (+57.0215 Da), variable Met oxidation (+15.9949 Da), variable N-terminal acetylation (+42.0106 Da), variable N-terminal dimethylation (+34.0631 Da), variable cyclization of N-terminal (i) glutamine (Gln → pyro-Glu; −17.0266 Da), (ii) glutamate (Glu → pyro-Glu; −18.0106 Da), and (iii) carbamidomethylated cysteine (pyro-cmC; −17.0266 Da). All identified PSMs were statistically evaluated using PeptideProphet[30] and then combined using iProphet,[31] as implemented in the Trans Proteomic Pipeline v4.8.0 PHILAE,[32] using a 1% false discovery rate (FDR) cutoff. Peptides were grouped and assigned to proteins using the ProteinProphet[33] module at a protein probability ≥0.95, corresponding to a FDR of 0.7%. If peptides matched multiple protein sequences, ProteinProphet determined the protein groups sharing the same set of peptides and identified for each protein group one protein entry as representative. Identified proteins were further annotated using the neXtProt release 2014−09−19,[2] the retrieve/ID mapping tool from UniProtKB,[34] and the knowledgebase TopFIND.[35−37]

### Bioinformatics Analyses

N-termini were defined according to their position within the corresponding protein sequence and their N-terminal amino acid modification. Classification as N-termini required: (i) a N-terminus carrying a heavy dimethyl, indicative of a free N-terminus in vivo as reductive dimethylation is performed in TAILS at the protein level and not at the peptide level as in shotgun proteomics, or a natural modification such as acetylation or N-terminal pyro-Glu from terminal glutamate (Glu → pyro-Glu), and (ii) a C-terminal arginine residue matching the specificity of trypsin as lysines are experimentally dimethylated prior to digest and thus not recognized. Peptides with blocked N-termini derived from N-terminal cyclization of glutamine (Gln → pyro-Glu) or carbamidomethylated cysteine (pyro-cmC), both known side reactions under the employed conditions, were used together with carryover tryptic peptides possessing free N-termini to improve protein identification only. Identified N-termini corresponding to proteins identified at ≤1.0% FDR were further annotated using the N-terminomics tool TopFINDER.[37] The classification of protein identifications into known and "missing" proteins was based on neXtProt classes PE1 and PE2−PE5, respectively (www.neXtProt.org, release 2014-09-19). Identified missing proteins were further validated by manually inspecting all corresponding 1159 PSMs on top of the statistical analysis by PeptideProphet and ProteinProphet,[30,33] leading to a final list of highest confidence missing protein entries. The mass spectrometry raw data and metadata have been deposited to ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository[38] with the PXD identifier <PXD002264>.

### Gene Ontology Analysis

Two orthogonal approaches were chosen for Gene Ontology (GO) analysis of pulp specific proteins. (i) GO annotations were binned into high-level GO Slim using GoTermFinder[39] and plotted as bar graphs comparing the respective cluster frequency with the global annotation frequency in the human proteome. (ii) A biological network analysis was performed using EnrichmentMap v2.1.0[40] within Cytoscape 3.2.1.[41] Nodes represent enriched Gene Ontology annotations as determined by BiNGO 3.0.3 using default parameters.[42] Gene sets were only included if they passed both $p$-value ($p \leq 0.001$) and $q$-value ($q \leq 0.05$) thresholds. Node size corresponds to the respective number of proteins in each category, and edges represent the association between terms, with edge thickness reflecting their overlap. Highly connected terms were grouped and annotated using WordCloud 3.0.1.[43]

## RESULTS

### The Dental Pulp Proteome and N-Terminome

We collected healthy pulp specimens from surgically extracted dental caries-free wisdom teeth from nine otherwise healthy young patients to map the dental pulp proteome and to simultaneously profile protein N-termini and in vivo proteolytic processing on a tissue-wide level. A total of 38 data sets, 22 TAILS and 16 preTAILS, were collected and analyzed using four different database search engines, namely Mascot,[26] X! Tandem,[27] MS-GF+,[28] and Comet.[29] Data were processed and combined at a FDR of ≤1.0% using PeptideProphet,[30] iProphet,[31] and ProteinProphet,[33] all bundled within the Trans-Proteomic Pipeline (TPP).[32]

First, we verified N-terminal enrichment (Figure 3A). In preTAILS only 9% of all identified spectra matched to N-terminally blocked peptides, which constituted over 94% of the postenrichment (TAILS) sample, reflecting a >10-fold N-terminal enrichment. A closer look at the identified blocked N-termini (Figure 3B) revealed that almost equal amounts of naturally acetylated (37%) and experimentally dimethylated (42%) were identified. Acetylation represents the most common intracellular N-terminal co- and post-translational modification with more than 9400 documented instances in the knowledgebase TopFIND,[37] whereas dimethylation indicates free in vivo N-termini, most of them originating from extracellular proteins and proteolytic processing. Together with pyroE from N-terminal glutamate (1%), acetylated and dimethylated peptides represent natural protein N-termini, as otherwise cyclized N-termini (20%) resulting from N-terminal glutamine or carbamidomethylated cysteine result from known side reactions on internal tryptic peptides and so are used for protein identification only. To double check mass accuracy, we analyzed the average ppm-offset of matched parental ions and found it to be 0.5877 ppm (2.75 ppm for their absolute values), with a standard deviation of 4.22; 84.5% of all identified PSMs were within ±5 ppm, and >95% within ±10 ppm.

Using four search engines and a FDR of ≤1.0% at PeptideProphet level, we identified 375 246 high-confidence peptide-spectrum matches (PSMs) (Figure 3C), corresponding to more than 21 000 modification-specific peptides and an average of 18 PSMs per identified peptide. 137 306 spectra (37%) were matched by all four search engines, and over 78%



DOI: 10.1021/acs.jproteome.5b00579
J. Proteome Res. 2015, 14, 3568−3582

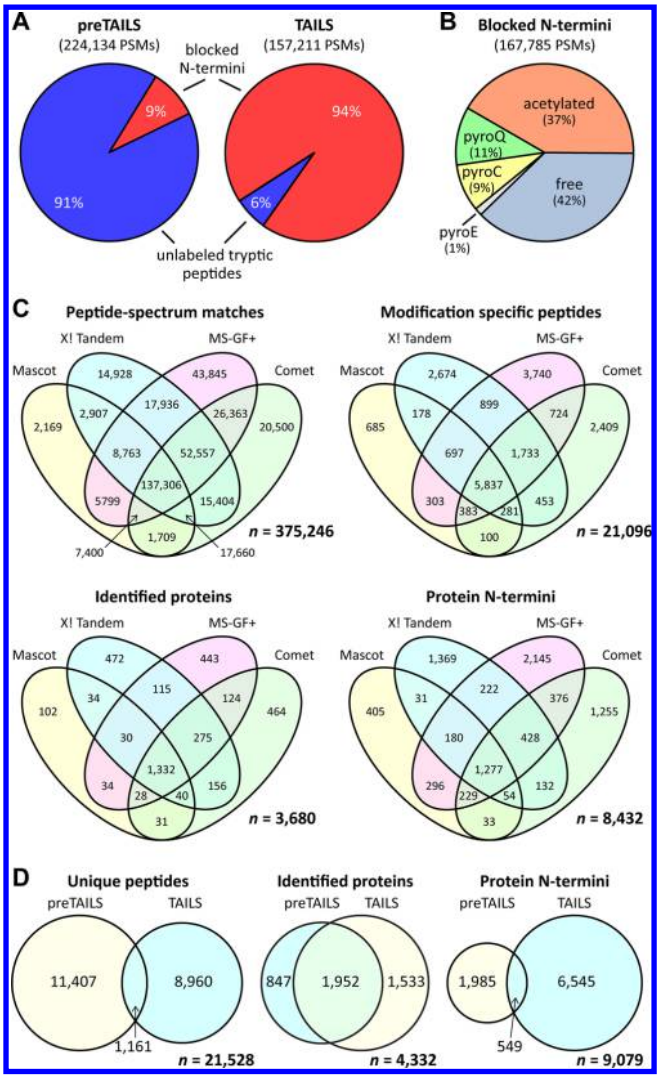



**Figure 3.** (A) Comparison of shotgun-like preTAILS versus TAILS N-terminomics peptide and protein features and identification characteristics. Distribution of peptide-spectrum matches (PSMs) identifying trypsin-generated internal peptides (blue) and blocked N-terminal peptides (red) are shown as pie charts. A greater than 10-fold enrichment for N-terminal peptides was calculated allowing in depth N-terminomics studies. (B) Unraveling the features of blocked N-termini. The vast majority of identified blocked N-termini are naturally acetylated (37%) or free (42%; i.e., experimentally dimethylated during TAILS). The remaining 21% originated from N-terminal cyclization, a known side reaction under the employed conditions for N-terminal glutamine (pyroQ) and carbamidomethylated cysteine (pyroC), but enzyme mediated in the case of N-terminal glutamate (pyroE). Only cyclized peptides from the natural N-terminus were included in the analyses as the remainder included bona fide neo-N termini, but also cyclized internal tryptic peptides. Nonetheless, such peptides can be used for improved protein identification or isoform assignment. (C) Comparison of peptide-spectrum matches, modification specific peptides, proteins, and protein N-termini identified by semispecific database searches using Mascot, X! Tandem, MS-GF+, and Comet (FDR ≤ 1% at PeptideProphet level). (D) Comparison of identified peptides, proteins, and protein N-termini by our two-pronged proteomics-N-terminomics approach. Even though preTAILS outnumbered TAILS in peptide identifications, TAILS identifies 20% more proteins, and nearly 4-times more protein N-termini (FDR ≤ 1% at iProphet level).

(293 804 PSMs) by at least two. Evaluating the individual search engines, MS-GF+ matched nearly 300 000 spectra (80% of all PSMs), followed by Comet (278 899; 74%), X! Tandem (267 461; 71%), and Mascot (183 713; 49%). A similar picture was obtained when comparing modification specific peptides: MS-GF+ identified nearly 70% of all identified unique peptides, followed by X! Tandem (60%) and Comet (56%), whereas on protein level (protein probability ≥0.95%), MS-GF+, X! Tandem, and Comet performed rather equally, and identified ∼2450 proteins and a total of 8432 protein N-termini. By combining all four search engines using iProphet, we identified at 1.0% FDR a total of 21 528 modification-specific peptides matching to nearly 400 000 spectra (Figure 3D). Even though preTAILS identified 25% more modification-specific peptides (12 568 vs 10 121), TAILS identified 25% more proteins (3485 vs 2799), and by far more natural protein N-termini (7094 vs 2534). In total, we identified 9079 high confidence protein N-termini: 3508 naturally acetylated, 217 cyclized (pyro-E from glutamate), and 5354 free in vivo N-termini at iProphet level.

As anticipated from previous reports, >70% of all termini were identified exclusively by TAILS, highlighting the strength of this N-terminomics approach, and 78% of all identified N-termini represented internal neo-N-termini, indicative of pervasive proteolytic processing even in healthy human tissues (Figure 4A). Strikingly, >25% of these were acetylated, documenting the importance of N-terminal acetylation as a post-translational modification besides its well-defined function as the recently described cotranslational modification by our group and Dr. K. Gevaert's group.[71] preTAILS contributed 1985 unique protein N-termini, which were not detected in TAILS, presumably due to sample loss after the negative enrichment step where low peptide concentrations or ultrafiltration can lead to losses of certain types of peptides (Figure 3D). In total, we identified by our combined shotgun-like (preTAILS) and N-terminomics (TAILS) approach, 4332 proteins at a ProteinProphet probability of ≥0.95, corresponding to a protein FDR of ≤0.7%. Thus, this study is the most extensive dental pulp proteomics study for any species to date. Spreadsheets containing all identified peptide spectrum matches for each individual search engine and of the combined analysis can be found at ProteomeXchange with the PXD identifier <PXD002264>.

## Mapping the Pulp N-Terminome

As a next step, we mapped all identified acetylated and dimethylated N-termini (≤1% FDR at iProphet level) on high confidence proteins (≤0.7% FDR at ProteinProphet level): 2674 modification specific N-terminally acetylated peptides were matched to 2128 unique N-termini in 1833 proteins, and 3964 N-terminally dimethylated peptides to 3518 unique N-termini and 2046 proteins, giving a total of 5646 protein termini used for further analysis (Figure 4A). The majority of the UniProt[34] annotated natural protein N-termini (1254 instances) represented position 1 and 2 of intact protein chains (341 and 771, respectively), whereas 11% corresponded to annotated cleavage sites of signal peptides (119) or prodomains (23). In total, we identified over 4300 internal protein N-termini, including 277 either starting with or immediately after methionine (156 and 121, respectively), hinting to potential alternative translation initiation sites as they showed a similar N-terminal methionine excision profile (see below). Using the web tool TopFINDER (http://clipserve.clip.ubc.ca/topfind/topfinder), 15% of these (42 instances) could be mapped to currently known alternative splicing (16) and alternative translation initiation (26) events, a


DOI: 10.1021/acs.jproteome.5b00579
J. Proteome Res. 2015, 14, 3568−3582

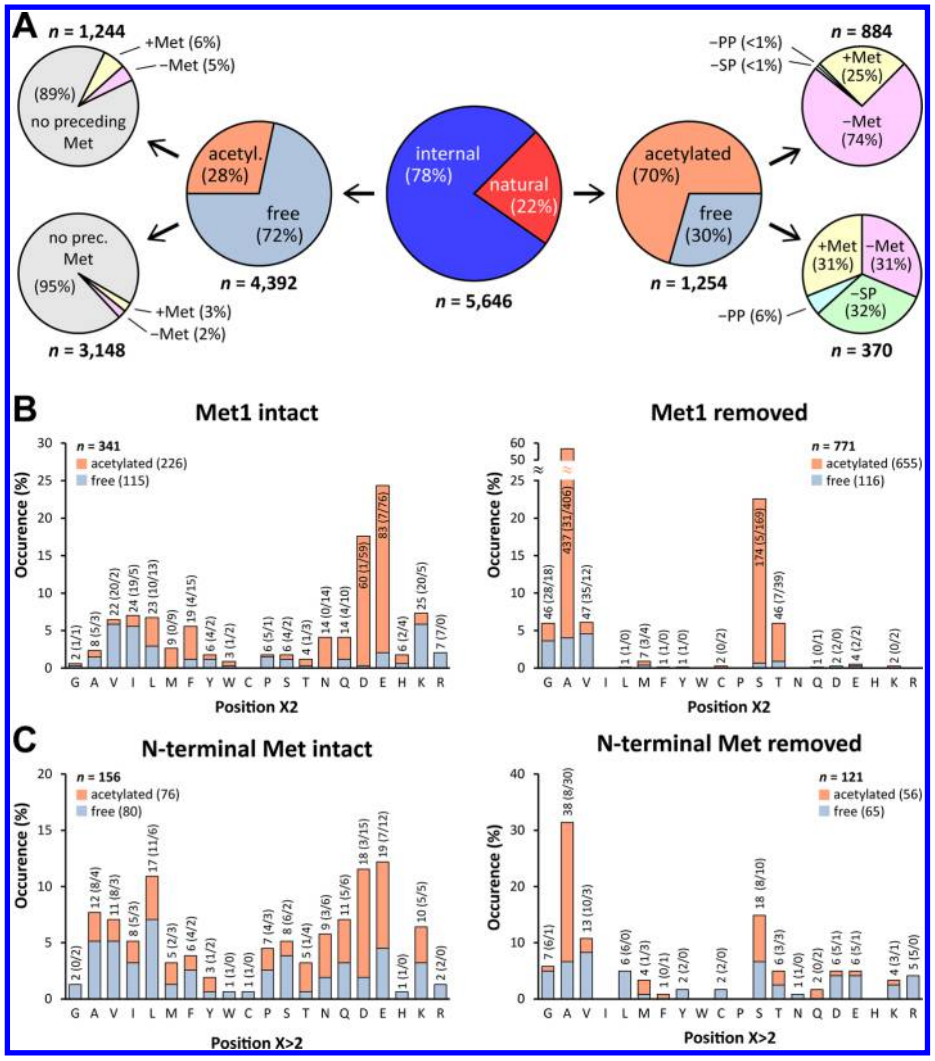



**Figure 4.** (A) In depth characterization of identified natural and internal protein N-termini and their acetylation status. A total of 5646 high confidence (FDR ≤ 1.0%) N-termini mapping to 4332 proteins were analyzed, with the distribution of protein N-termini classified as natural (i.e., as translated or after classical protein maturation) or internal (i.e., derived from proteolytic processing) shown. Fractions of N$\alpha$-acetylated and free (dimethylated during TAILS) N-termini are indicated. Frequency distribution of protein maturation events such as classical N-terminal methionine excision, signal-peptide cleavage, and prodomain removal are highlighted. (B) Characterization of N-terminal methionine excision and subsequent acetylation of natural protein N-termini (i.e., representing position 1 or 2 of the UniProt annotated protein chain). In total, 341 N-termini possessing an N-terminal methionine (top) and 771 with initiator Met removed (bottom) were identified and analyzed. Fractions of identified N$\alpha$-acetylated versus free N-termini are indicated, together with frequency distributions of individual amino acids. (C) Characterization of N-terminal methionine excision and subsequent acetylation of internal protein neo-N-termini (i.e., not representing position 1 or 2 of the UniProt annotated protein chain). In total, 121 N-termini possessing an N-terminal methionine (top) and 156 with initiator Met removed (bottom) were identified and analyzed. Fractions of identified N$\alpha$-acetylated vs. free N-termini are indicated, and frequency distributions of individual amino acids are shown.

number supposed to increase in the near future with more proteogenomics data being published on alternative translation initiation.[44,45] It was interesting to capture proteins in the act of protein synthesis by TAILS; 10 extracellular proteins were identified with their signal peptides present or only partly processed (Table S1), including the protein Kazal-type serine protease inhibitor domain-containing protein 1, also known as bone and odontoblast-expressed protein 1, indicating their localization within the secretory pathway.

Consistent with the human platelet N-terminome,[7] we observed distinct preferences for N-terminal methionine excision (Figure 4B and 4C; Table S2). Methionine followed by charged (Asp, Glu, Lys, Arg), large hydrophobic (Ile, Leu, Phe, Trp, Tyr) or large polar (Asn, Gln) residues was virtually never removed (4%), whereas N-terminal methionine excision was consistently observed (97%) when methionine was followed by small hydrophobic (Gly, Ala) or small polar residues (Ser, Thr). For example, we identified in total 180 protein N-termini with Ser in position 2. Of these, in 174 cases the initiator methionine was removed confirming the high fidelity of methionine amino peptidase 1, whereas in 83 out of 87 protein N-termini with Glu in position 2 the initiator methionine was unprocessed. Only Val and Met showed an intermediate profile with an N-terminal methionine occurrence of 67% and 44%, respectively.

Interestingly, Ala, Ser, and Thr were nearly exclusively acetylated after N-terminal methionine excision (93%), whereas for Gly more than 60% free N-termini were identified, a value similar to Val (75%). N-terminal Met was frequently acetylated when followed by Asp, Glu, Asn, Met, or Phe (94%), but remained unblocked in case of Val (91%), Ile (79%), and Lys (80%) in position 2; Leu and Gln induced intermediate Met acetylation profiles (57% and 71%, respectively). Several



DOI: 10.1021/acs.jproteome.5b00579
J. Proteome Res. 2015, 14, 3568−3582

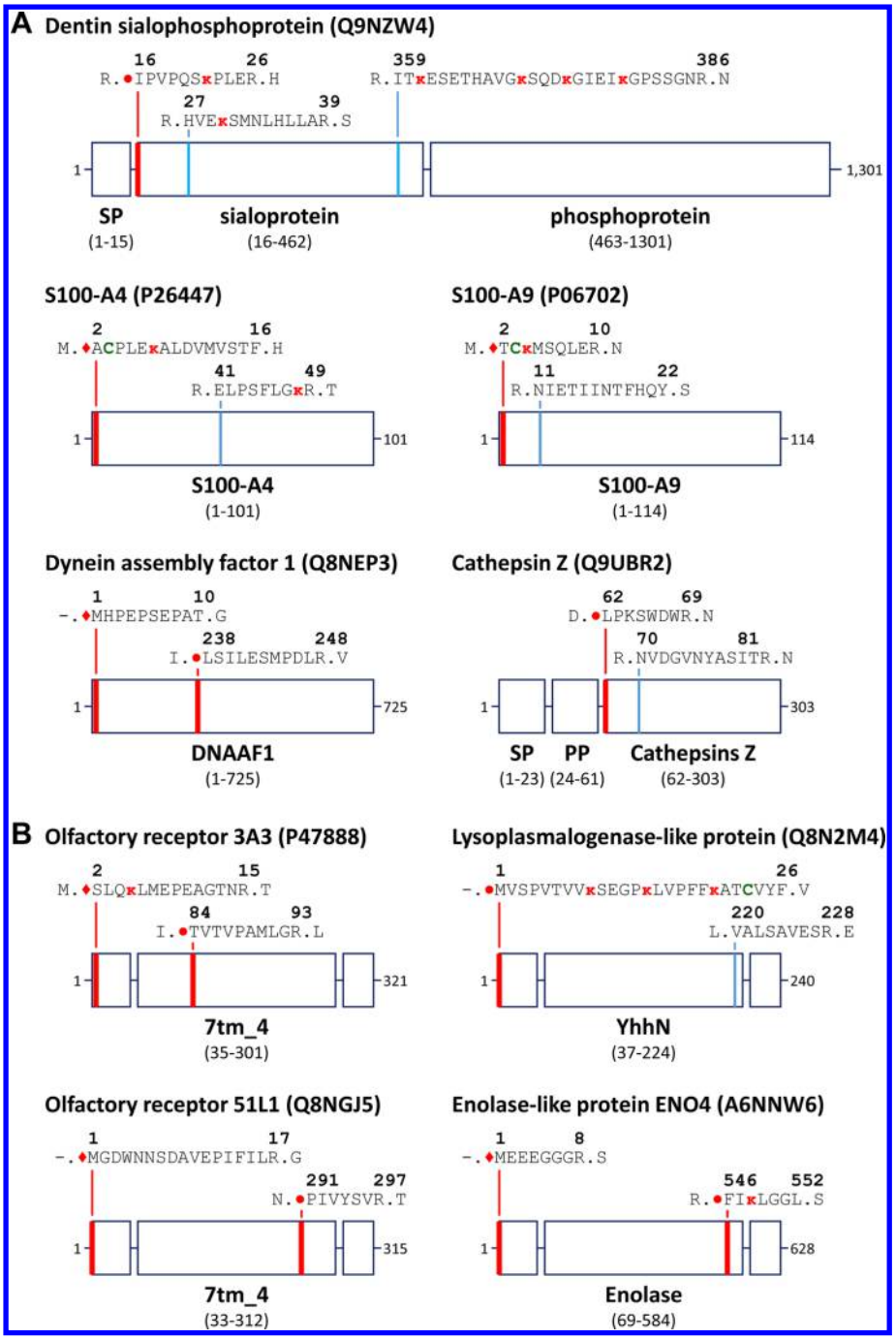



**Figure 5.** Examples of proteins identified in human dental pulp with the identified N-termini (TAILS; red) and internal peptides (preTAILS; blue) mapped. (A) Dentin sialophosphoprotein (DSPP), two representatives of the S100A protein family, dynein assembly factor 1 (DNAAF1), and cathepsin Z are shown. Dentin sialophosphoprotein was identified in its secreted form (i.e., lacking its signal peptide) by three high confidence peptides including its mature N-terminus starting at position 16. Notably, both S100A proteins were identified after N-terminal methionine excision and with acetylated Ala and Thr at their N-terminus, respectively, in perfect accordance with the identified N-terminal methionine excision and acetylation profiles (see Figure 4). DNAAF1 is one of the 148 proteins identified related to cilia and/or cell projection processes and therefore are likely located to the odontoblast dentin process. Cathepsin Z was exclusively identified in its secreted active form, i.e., after signal peptide and prodomain removal and hence represents the activated protease. (B) Four proteins so far classified as "missing" but identified in this study by TAILS N-terminomics in dental human pulp. Two olfactory receptors, namely 3A3 and 51L1, lysoplasmalogenase-like protein TMEM86A, and enolase-like protein ENO4 are depicted. All four were identified via their mature N-terminus at position 1 or 2, and one more peptide representing either a neo-N-terminus or an internal peptide. Sequence positions and amino acid modifications are indicated: (red diamond) acetylated N-terminus, (red circle) free N-terminus (dimethylated in TAILS), (red K) dimethylated lysine, (green C) carbamidomethylated cysteine.

examples of identified proteins and their proteotypic protein N-termini are shown in Figure 5, including the major secretory product of odontoblasts, dentin sialophosphoprotein, which is crucial for proper tooth formation.[46]

Of note, we identified N-termini of five cysteine cathepsins (cathepsins B, C, H, K, and Z) in their activated form, i.e., starting exactly with the UniProt annotated residue after prodomain removal, as shown for cathepsin Z in Figure 5, and three of these



DOI: 10.1021/acs.jproteome.5b00579
J. Proteome Res. 2015, 14, 3568−3582

can be linked to the dentin-pulp complex via published reports.[47−50] Cathepsins were initially thought to be confined to lysosomes, but over the last 10 years a broad spectrum of extra-lysosomal functions was uncovered. For example, cathepsin K represents the most potent mammalian collagenase and has an essential role in bone resorption;[47] as dentin resembles bone and odontoblasts respond to tooth injury by forming tertiary dentin, an involvement of cathepsins K is easily supposed.[48] Furthermore, cathepsin C deficiency leads to Papillon-Lefèvre Syndrome, a severe form of juvenile periodontitis,[50] and cathepsin B was detected in gingival crevicular fluid during orthodontic tooth movement.[49] and has been recently been identified as a potent processor of several chemokines.[51] Additionally, we identified the cysteine protease inhibitors cystatin-B and cystatin-D by their N-terminal peptides. The intracellular cystatin B was found with an intact and acetylated initiator methionine, whereas the N-terminus of the extracellular cystatin-D inhibitor was unblocked after cotranslational methionine removal, but prior to signal-peptide cleavage, indicating its identification and transient localization in the endoplasmic reticulum. Finally in the protease realm we identified 3 matrix metalloproteinases (MMP-9, -23, and -25) and 5 ADAMTS proteases (ADAM-TS 2, 7, 12, 19, and 20) involved in signal processing and extracellular matrix remodeling.[52,53]

## Comparison with Previous Proteomics Data Sets Identifies Unique Pulp Proteins

The human dental pulp proteins identified in this study included 70% of the 373 proteins reported by Pääkkönen et al. 2005[20] and Eckhardt et al. 2014,[21] who used 2-dimensional gel electrophoresis followed by tandem mass spectrometry (Figure 6A). Furthermore, we covered 55% of the published dentin proteome,[54] due to the presence and coanalysis of the single cell deep odontoblast layer in our pulps, pointing to the remarkable sensitivity of TAILS in enriching for even low abundant N termini (Figure 2B). The pulp is expected to contain fibroblast and plasma proteins in addition to proteins derived from blood cells. To further identify pulp-enriched or pulp-specific proteins, we compared our protein list with the published proteomes of erythrocytes,[6] platelets,[7] plasma,[55] and human skin fibroblasts.[56,57] The resulting 1777 pulp-enriched proteins were used for gene ontology analysis using (i) a functional enrichment map using Cytoscape (Figure 6B) and (ii) GoTermFinder,[39] binning all annotations into high-level GO Slim terms (Figure 7). 1675 proteins were successfully mapped and grouped into 53 parental terms, of which all with more than 50 evidence were plotted and compared with the global human proteome. 1199 UniProt entries matched the GO term "intracellular" (72%), 316 "extracellular region" (19%), and 173 (10%) both GO terms. 57 entries were annotated as "proteinaceous extracellular matrix" (3.4%), including 2 heparan sulfate proteoglycans (glypican-4 and -6), 5 laminin subunits (alpha-1, alpha-3, beta-4, and gamma-2), 3 members of the Wnt family (Wnt-8b, 10a, and 11), and 9 collagens, including collagen type XXVII alpha-1, which plays an important role during the calcification of cartilage and its transition to bone[58] and hence we would anticipate in dentin also. Furthermore, we identified the noncollagenous extracellular matrix polyprotein dentin sialophosphoprotein, corroborating the presence of odontoblasts in our pulp preparations. Dentin sialophosphoprotein is secreted by odontoblasts and cleaved by astacin metalloproteases into dentin sialoprotein and dentin

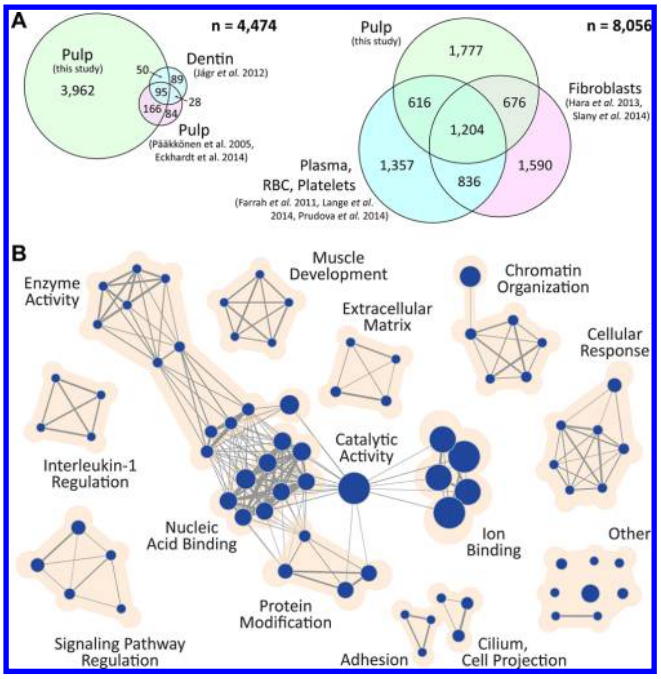



**Figure 6.** (A) Comparison of the human dental pulp proteome with previously published GeLC−MS/MS based proteomics studies of human dentin[54] and pulp[21] (left), and with human plasma,[55] erythrocytes,[6] platelets[7] and fibroblasts[56,57] (right). (B) Biological network analysis of 1777 proteins identified in human dental pulp, but not in plasma,[55] erythrocytes,[6] platelets,[7] or fibroblasts.[56,57] Nodes represents enriched biological process, molecular function and cellular compartment terms from Gene Ontology. Node size corresponds to the respective number of proteins in each category, and edges represent the association between terms, with edge thickness reflecting their overlap.

phosphoprotein, which in turn regulate the initiation and maturation of dentin mineralization, respectively.[59,60]

Enzymatic processes were well represented with 257 of 1777 proteins annotated with "hydrolase activity" (15%), 87 with "kinase activity" (5%), and a total of 549 with the more generic term "catalytic activity" (33%). Perusing the annotated biological processes 1350 (81%), 1086 (65%), and 955 (57%) entries reflected the terms "cellular process", "metabolic process", and "regulation of biological process", respectively, and for the top 10 entries, an average enrichment of 15% was observed when compared with the whole human proteome GO-term frequency. An alternative visualization using a gene ontology enrichment map of over-represented categories is depicted in Figure 6B.

## Chromosome Annotation

To support the C-HPP initiative we mapped the identified proteins to their encoding chromosomes (Figure 8) finding a relatively even representation of all chromosomes (average coverage of 21%), with the notable exception of the Y-chromosome (6.4%) and mitochondrial genes (7.1%). Not surprisingly, the lowest coverage for all other chromosomes was obtained for chromosome 21 (15.1%), which is known to contain many proteins with few mass-spectrometry-compatible tryptic peptides.[61] Of note, 20% and 35% of all identified proteins were identified only in preTAILS and TAILS analyses, respectively, and 45% were found in both, illustrating the orthogonality and hence power of these two complementary methods. Furthermore, by our combined proteomics approach, we identified 174 candidate missing proteins (PE levels 2−4; neXtProt release 2014−09−19) at a protein FDR of ≤0.7%, 41 by preTAILS, 63


DOI: 10.1021/acs.jproteome.5b00579
J. Proteome Res. 2015, 14, 3568−3582

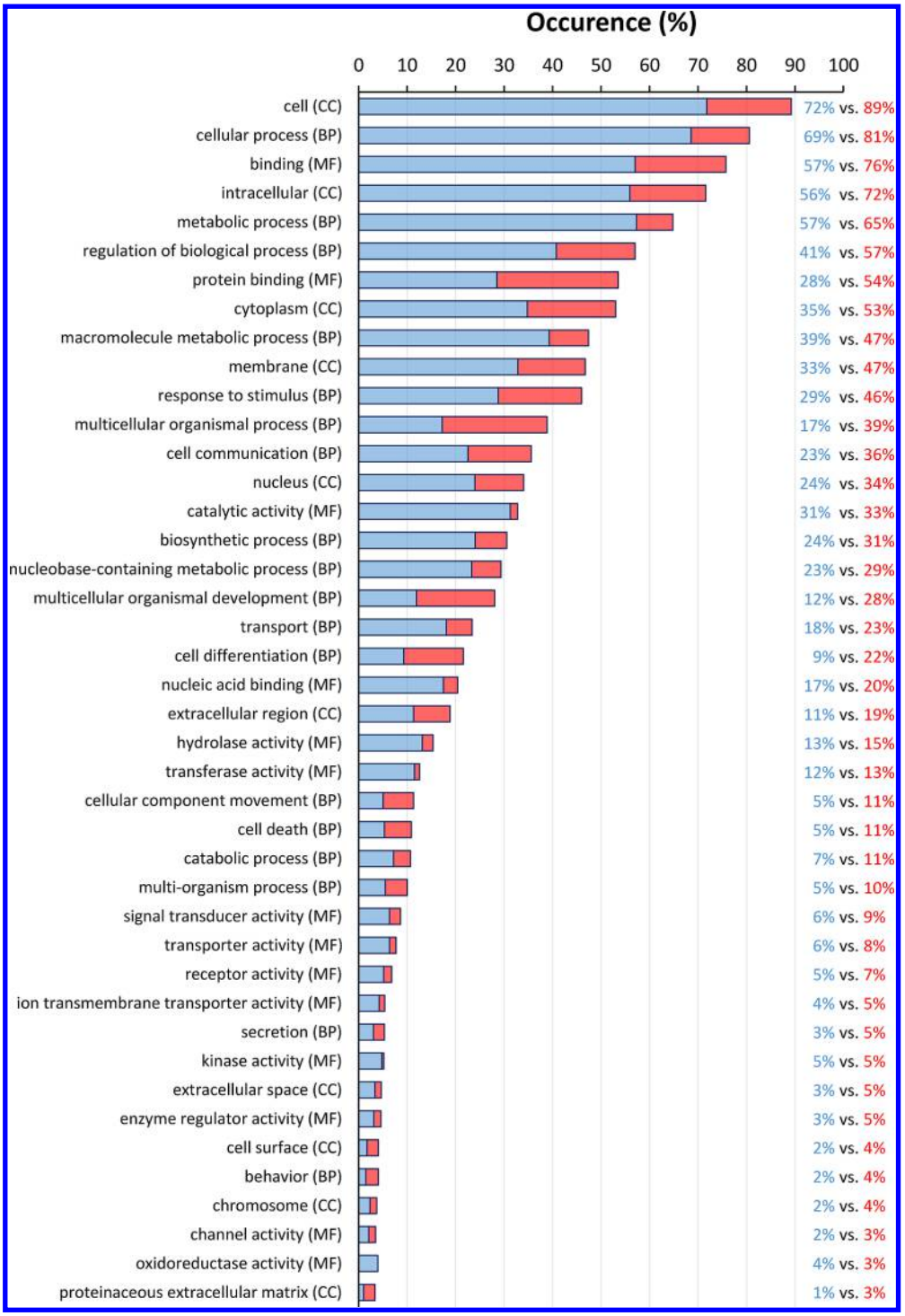



**Figure 7.** Gene ontology analysis of 1777 proteins identified in human dental pulp but not in plasma,[55] erythrocytes,[6] platelets,[7] or fibroblasts.[56,57] GO Slim terms with more than 50 instances are depicted and enrichment over the human genome is indicated in red. The respective whole genome and cluster frequencies within the 1777 proteins are depicted on the right. Of note, 148 entries annotated with "cell projection" and "cilium" (compare Figure 6B) are condensed with others in the parental GO Slim category "cell".

by both, and 70 were exclusively identified by TAILS. Other than for the Y-chromosome and mitochondrial DNA, we identified for every chromosome at least 1, and on average 7 candidate missing proteins; 96% corresponded to protein existence level 2 (evidence on transcriptome level), and seven derived from PE levels 3 (inferred from homology) and PE4 (predicted) categories (Tables S3 and S4), of which three have preliminary antibody detection results in the Human Protein Atlas.[62] We further evaluated all identified missing proteins based on their PSMs and used a credit rating-inspired nomenclature for a quality-assessed candidate missing proteins list (Table S5). In addition to a mandatory peptide identification FDR ≤ 0.01 and a ProteinProphet protein probability ≥0.95, approximately half of the candidate missing proteins (81 entries) had at least 2 PSMs with a PeptideProphet probability of ≥0.90 and various numbers of PSMs with probabilities ≥0.50 (but ≤0.90), and thus qualified



DOI: 10.1021/acs.jproteome.5b00579
J. Proteome Res. 2015, 14, 3568−3582

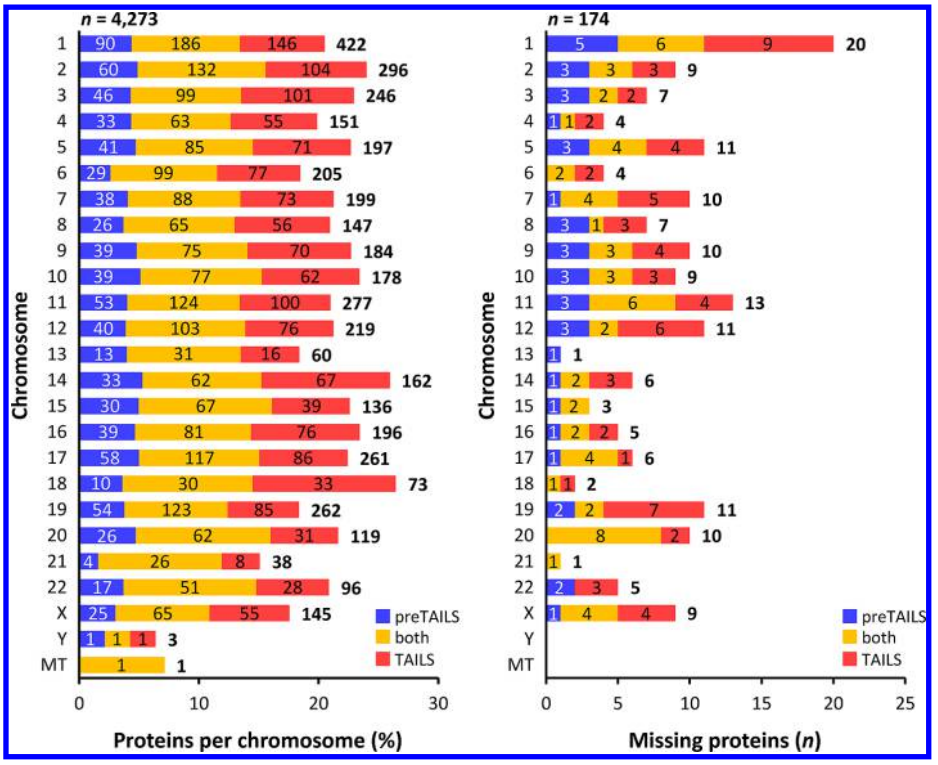



**Figure 8.** Chromosomal distribution of the genes encoding identified proteins displayed as percentage of total protein coding genes per chromosome (left). On the right: Distribution of "classified as missing"-proteins per chromosome. We identified 7.5 candidate missing proteins per autosome and 5 for the X-chromosome. Blue: proteins identified by preTAILS; orange: proteins identified by preTAILS and TAILS; red: proteins identified solely by TAILS.

for an A-rating; 4 of the 7 identified olfactory receptors fell into this category. The enolase-like protein ENO4, which is depicted in Figure 5B, was, e.g., identified by 14 PSMs with a probability of ≥0.90 and as for two of the olfactory receptors qualified for a triple-A rating. Nonetheless, as the C-HPP initiative expects extraordinary evidence for extraordinary claims,[63,64] the aforementioned statistics are not satisfactory on their own for a final missing protein list and are ideally substantiated by synthetic-peptide reference spectra. Therefore, we manually evaluated all 1159 PSMs pointing toward missing-proteins, checked all peptide sequences for Ile/Leu ambiguities against SwissProt, and removed candidate proteins corroborated only by a 7 amino acid PSM, before assembling a final, highly stringent candidate missing-protein list containing 17 of the previous 174 entries (Table S6). We include this step-by-step breakdown in the manuscript and Supplementary Tables to render the process transparent and to allow the reader to assess the quality of the individual identifications. Nevertheless, an additional level of validation of these candidate identifications by comparison of matched spectra with the spectra of corresponding synthetic peptides will be required to fulfill HPP-criteria, but these proteins are included here to point researchers toward these high confidence candidates in similar or different tissues. PSMs identifying candidate missing proteins were uploaded together with the mass spectrometry raw data to ProteomeXchange.

## DISCUSSION

Overall, we present the first proteome-wide study of protein N-termini and proteolytic processing from human dental pulp, revealing a remarkable 78% of termini being distinct from annotated original termini, and report over 4300 identified proteins, as well as 174 lacking evidence on the protein level (neXtProt level PE2−4, release 2014−09−19), 104 by preTAILS and 133 by TAILS thereby validating the approach of using unusual tissues and unusual peptides to seek missing protein candidates. Whereas the 70 proteins identified solely by TAILS highlights the strength of TAILS N-terminomics and semi-specific N-terminal peptides to complement conventional shot-gun proteomics, proteins identified by preTAILS reflect the exceptional value of using unusual low abundance human tissues as proteome sources to expose and identify missing proteins. Importantly, as the C-HPP set standards higher than the accepted statistics-based peptide and protein probabilities and FDRs for the ultimate identification of missing proteins, we manually inspected all respective PSMs for their fragmentation patterns and further checked matching peptides for Ile/Leu-ambiguities, giving a list of 17 former missing-protein candidates for the C-HPP. Nonetheless, like all approaches, one of the highest standard of validation of missing proteins will be from future comparative analysis of the spectra of synthetic peptides corresponding to the identified sequence with those PSMs obtained from the tissues.[63,64]

Our analysis is by far the most comprehensive reported for human dental pulp, greatly exceeding the two previous 2-dimensional electrophoresis studies. The combined use of four different database search engines proved to be highly beneficial as each algorithm identified between 685 to 3740 modification specific peptides that were missed by the other three. As we found before, MS-GF+ was the most effective in comprehensively identifying semitryptic peptides.[6] Thus, >9000 protein N-termini from nearly 400 000 peptide-spectrum matches at a FDR of ≤1.0% were identified by this two-pronged proteomics approach.



DOI: 10.1021/acs.jproteome.5b00579
J. Proteome Res. 2015, 14, 3568−3582

Interestingly, we identified eight missing protein candidates belonging to the chemosensory receptor gene repertoires: seven olfactory receptors and one taste receptor type 2 (Tables S3 and S4). This identifications is not surprising, as homologues have been shown to be ectopically expressed in various tissues, where they presumably act as sensors for the microenvironment or sentinels of innate immunity.[65,66] However, we posit that olfactory receptors may form the molecular basis for hyphenating with the external environment by transducing osmotic, pH and salt changes that can adversely affect dental health. Notably, we identified 5 missing proteins that are associated with cilia, i.e. 2 associated with cilia and 3 ciliary dyneins. Thus, the cilia most likely sense and transmit external stimuli such as mechanical and possibly thermal stress to the odontoblast layer, reporting also dental caries onset.

We also identified 17 serine protease inhibitors including the key regulator of the complement cascade, C1-inhibitor (serpin G1),[9] and could map nearly the whole complement pathway including the complement control proteins factor H and factor I. We identified (i) all subunits of the C1-complex (C1q, C1r, and C 1s), (ii) the central elements C2, C3, C4 and C5, (iii) the N-termini of the anaphylatoxins C3a and C4a (C5a was most likely missed as its N-terminal peptide comprises 35 amino acids under the applied conditions), and (iv) with exception of C7, all components of the terminal membrane attack complex (MAC; C5b, C6, C7, C8, and C9), highlighting the comprehensive nature of our study and the sensitivity of TAILS.

We hypothesized that the analysis of human dental pulp, an unusual tissue, with a nonshotgun proteomics technique would allow for the identification of missing proteins alongside a myriad of undescribed natural protein N-termini that are mechanistically informative in characterizing protein proteforms. This hypothesis is supported by our previous studies on human erythrocytes[6] and platelets[7] that validate the suitability of TAILS N-terminomics to identify proteins spanning a wide dynamic range of abundance of 5 orders of magnitude.[6] We analyzed 1254 natural protein N-termini, 341 with the initiator methionine present and 771 removed, resulting in a robust assessment of the N-terminal methionine excision and acetylation profiles of the individual amino acids. When Ala, Gly, Ser or Thr were present in position 2, N-terminal methionine excision happened nearly without exception (>92%), and was followed by an amino acid residue-dependent frequency of N$\alpha$-acetylation: whereas Ala and Ser were predominantly acetylated (>92%), N-terminal glycine remained frequently unblocked (61%). On the contrary, the presence of glutamate or aspartate in position 2 virtually abolished N-terminal methionine excision (<5%), but induced $\alpha$-acetylation (>92%), in perfect agreement with previous studies and the N-end rule.[6,7,9,10,67] To our surprise we also identified 10 proteins with unprocessed signal peptides, indicating their presence in the secretory pathway and "caught in the act" of protein synthesis, including two extracellular protease inhibitors.

In total, we found 8 human pulp-orientated proteomics studies in the literature, but only 2 analyzed pulp whereas 6 analyzed ex vivo cultured cells and none used a gel-free proteomics approach, considered optimal for whole proteome analysis.[68] The most comprehensive analysis to date was by Pääkkönen et al. 2005, who identified 96 proteins in pulp from healthy and carious human teeth,[20] and Morsczeck et al. 2009,[69] who reported 94 differently regulated proteins during osteogenic differentiation of dental follicle precursor cells in vitro. Only one other study aimed to annotate the human pulp proteome: Eckhardt et al. 2014[21] identified 342 pulp proteins, including 37 not identified in human dentin or plasma. In comparison, by our comprehensive gel-free approach, we identified 4332 proteins, of which 1777 were not reported in previous N-terminomics studies on platelets[7] and erythrocytes,[6] or shotgun studies on fibroblasts[56,57] and plasma,[55] which could reasonably be assumed to be present in pulp. Furthermore, we covered 55% of the published dentin proteome due to the presence of odontoblasts, and 70% of the previously published pulp proteome. But even though we identified 10× more proteins than previously reported by gel based MS/MS methods,[20,21] 112 pulp proteins were not reported as being pulp, but these were identified only with a 5% FDR. We also identified several dentin-specific proteins originating from the odontoblast layer (e.g., dentin sialophosphoprotein, dermatopontin, chondrocalcin), which comprises only a minor part of the whole pulp, highlighting the sensitivity of TAILS to find low abundant proteins from low abundant cells.

Building on previous N-terminomics studies on specialized human cells,[6,7] our present study establishes a general TAILS N-terminomics workflow suitable for the in-depth characterization of human protein N-termini in tissues, which will allow the study of other tissues and cells, and various disease states for the HPP. Our strategy of (i) using a rare tissue together with (ii) proteome simplification by positional proteomics and (iii) targeting semitryptic peptides of the natural N-terminus or those generated by proteolytic processing in vivo, which exhibit altered mass spectrometric properties compared to their fully tryptic counterparts, proved to be highly fruitful in complementing conventional shot-gun proteomics approaches. In addition, this overall strategy provides valuable information for the ongoing efforts of the Human Proteome Project to complete the human proteome map. Enriching for semitryptic peptides that may be more amenable to identification than their spanning parent tryptic peptide due to altered $m/z$, ionization, and fragmentation properties, TAILS N-terminomics allows the identification of a subset of proteins which are otherwise inaccessible to shot-gun proteomics.[6] This study also provides an unbiased system-wide surveillance of global proteolytic processing in a healthy human tissue for the first time, complementing previous analysis in murine tissues[9,70] and highlighting the importance of mechanistic informative protein N-termini in understanding biological processes.

## ■ ASSOCIATED CONTENT

### Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jproteome.5b00579.

Table S1: Proteins identified with an unprocessed signal peptide. Table S2: (A) N-terminal methionine excision and acetylation preferences of original protein termini (Position 1 or 2). (B) N-terminal methionine excision and acetylation preferences of internal protein termini (any position but 1 and 2). Table S3: Identified "missing protein"-candidates using TAILS N-terminomics. Table S4: Peptide spectrum matches (PSMs) of identified "missing protein"-candidates. Table S5: Ranking of "missing protein"-candidates based on their PSM PeptideProphet-Probabilities. Table S6: Manually evaluated "missing protein"-candidates for the C-HPP. (XLSX)



DOI: 10.1021/acs.jproteome.5b00579
J. Proteome Res. 2015, 14, 3568−3582

## AUTHOR INFORMATION

### Corresponding Author

*Phone: +1-604-822-2958. Fax: +1-604-822-7742. E-mail: chris.overall@ubc.ca.

### Author Contributions

U.E. and C.M.O designed the research. U.E., G.M, S.R.A., and G.T. performed experiments, and I.M. conducted wisdom tooth extractions. U.E. analyzed and interpreted data together with G.M. and G.T., U.E. and G.M. prepared figures and wrote the initial manuscript. C.M.O. supervised the project, helped with data interpretation and manuscript revision, and provided grant support. All authors were involved in critically revising the manuscript, and approved the final version for publication.

### Notes

The authors declare no competing financial interest.

## ACKNOWLEDGMENTS

The authors thank all current members of the Overall Lab for inspiring discussions, feedback, and support, especially G. Butler for outstanding editing assistance, Jason Rogalski from the University of British Columbia Proteomics Core Facility for excellent LC−MS/MS measurements, David Shteynberg and Luis Mendoza from the Institute for Systems Biology (Seattle, Washington, USA) for coding support within TPP and the add on software CLIPPER, and Dr. Ian Matthew's clinical staff for consenting patients and their outstanding on-site support at the dental clinic. U.E. was supported by a postdoctoral fellowship from MSFHR, G.M. was cofunded by a UBC Centre for Blood Research Internal Collaborative Training Award, and C.M.O. by Canada Research Chair in Proteinase Proteomics and Systems Biology. A project grant from the Canadian Institutes of Health Research (MOP-133632) as well as infrastructure grants from the Michael Smith Research Foundation for Health Research (MSFHR) and the Canada Foundations for Innovation (CFI) supported the research. None of the funders had a role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

## REFERENCES

(1) Legrain, P.; Aebersold, R.; Archakov, A.; Bairoch, A.; Bala, K.; Beretta, L.; Bergeron, J.; Borchers, C. H.; Corthals, G. L.; Costello, C. E.; et al. The human proteome project: current state and future direction. *Mol. Cell. Proteomics* **2011**, *10* (7), M111.009993.

(2) Gaudet, P.; Argoud-Puy, G.; Cusin, I.; Duek, P.; Evalet, O.; Gateau, A.; Gleizes, A.; Pereira, M.; Zahn-Zabal, M.; Zwahlen, C.; et al. neXtProt: organizing protein knowledge in the context of human proteome projects. *J. Proteome Res.* **2013**, *12* (1), 293−298.

(3) Paik, Y.-K.; Omenn, G. S.; Uhlen, M.; Hanash, S.; Marko-Varga, G.; Aebersold, R.; Bairoch, A.; Yamamoto, T.; Legrain, P.; Lee, H.-J.; et al. Standard Guidelines for the Chromosome-Centric Human Proteome Project. *J. Proteome Res.* **2012**, *11* (4), 2005−2013.

(4) Gaudet, P.; Argoud-Puy, G.; Cusin, I.; Duek, P.; Evalet, O.; Gateau, A.; Gleizes, A.; Pereira, M.; Zahn-Zabal, M.; Zwahlen, C.; et al. neXtProt: organizing protein knowledge in the context of human proteome projects. *J. Proteome Res.* **2013**, *12* (1), 293−298.

(5) Lane, L.; Bairoch, A.; Beavis, R. C.; Deutsch, E. W.; Gaudet, P.; Lundberg, E.; Omenn, G. S. Metrics for the Human Proteome Project 2013−2014 and strategies for finding missing proteins. *J. Proteome Res.* **2014**, *13* (1), 15−20.

(6) Lange, P. F.; Huesgen, P. F.; Nguyen, K.; Overall, C. M. Annotating N termini for the human proteome project: N termini and Nα-acetylation status differentiate stable cleaved protein species from degradation remnants in the human erythrocyte proteome. *J. Proteome Res.* **2014**, *13* (4), 2028−2044.

(7) Prudova, A.; Serrano, K.; Eckhard, U.; Fortelny, N.; Devine, D. V.; Overall, C. M. TAILS N-terminomics of human platelets reveals pervasive metalloproteinase-dependent proteolytic processing in storage. *Blood* **2014**, *124* (26), e49−e60.

(8) Kleifeld, O.; Doucet, A.; auf dem Keller, U.; Prudova, A.; Schilling, O.; Kainthan, R. K.; Starr, A. E.; Foster, L. J.; Kizhakkedathu, J. N.; Overall, C. M. Isotopic labeling of terminal amines in complex samples identifies protein N-termini and protease cleavage products. *Nat. Biotechnol.* **2010**, *28* (3), 281−288.

(9) Auf dem Keller, U.; Prudova, A.; Eckhard, U.; Fingleton, B.; Overall, C. M. Systems-level analysis of proteolytic events in increased vascular permeability and complement activation in skin inflammation. *Sci. Signaling* **2013**, *6* (258), rs2.

(10) Marino, G.; Eckhard, U.; Overall, C. M. Protein Termini and Their Modifications Revealed by Positional Proteomics. *ACS Chem. Biol.* **2015**, DOI: 10.1021/acschembio.5b00189.

(11) Rogers, L. D.; Overall, C. M. Proteolytic post-translational modification of proteins: proteomic tools and methodology. *Mol. Cell. Proteomics* **2013**, *12* (12), 3532−3542.

(12) Butler, G. S.; Overall, C. M. Matrix metalloproteinase processing of signaling molecules to regulate inflammation. *Periodontol. 2000* **2013**, *63* (1), 123−148.

(13) Couve, E.; Osorio, R.; Schmachtenberg, O. The amazing odontoblast: activity, autophagy, and aging. *J. Dent. Res.* **2013**, *92* (9), 765−772.

(14) Staquet, M.-J.; Carrouel, F.; Keller, J.-F.; Baudouin, C.; Msika, P.; Bleicher, F.; Kufer, T. A.; Farges, J.-C. Pattern-recognition receptors in pulp defense. *Adv. Dent. Res.* **2011**, *23* (3), 296−301.

(15) Liu, H.; Gronthos, S.; Shi, S. Dental pulp stem cells. *Methods Enzymol.* **2006**, *419*, 99−113.

(16) Jontell, M.; Okiji, T.; Dahlgren, U.; Bergenholtz, G. Immune defense mechanisms of the dental pulp. *Crit. Rev. Oral Biol. Med.* **1998**, *9* (2), 179−200.

(17) Trope, M. Regenerative potential of dental pulp. *J. Endod.* **2008**, *34* (7 Suppl), S13−S17.

(18) Pääkkönen, V.; Tjäderhane, L. High-throughput gene and protein expression analysis in pulp biologic research: review. *J. Endod.* **2010**, *36* (2), 179−189.

(19) Jágr, M.; Eckhardt, A.; Pataridis, S.; Broukal, Z.; Dušková, J.; Mikšík, I. Proteomics of human teeth and saliva. *Physiol. Res.* 2014, *63* (Suppl 1), S141−S154; PMID: 24564654.

(20) Pääkkönen, V.; Ohlmeier, S.; Bergmann, U.; Larmas, M.; Salo, T.; Tjäderhane, L. Analysis of gene and protein expression in healthy and carious tooth pulp with cDNA microarray and two-dimensional gel electrophoresis. *Eur. J. Oral Sci.* **2005**, *113* (5), 369−379.

(21) Eckhardt, A.; Jágr, M.; Pataridis, S.; Mikšík, I. Proteomic analysis of human tooth pulp: proteomics of human tooth. *J. Endod.* **2014**, *40* (12), 1961−1966.

(22) Kleifeld, O.; Doucet, A.; Prudova, A.; auf dem Keller, U.; Gioia, M.; Kizhakkedathu, J. N.; Overall, C. M. Identifying and quantifying proteolytic events and the natural N terminome by terminal amine isotopic labeling of substrates. *Nat. Protoc.* **2011**, *6* (10), 1578−1611.

(23) Wessel, D.; Flügge, U. I. A method for the quantitative recovery of protein in dilute solution in the presence of detergents and lipids. *Anal. Biochem.* **1984**, *138* (1), 141−143.

(24) Rappsilber, J.; Mann, M.; Ishihama, Y. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat. Protoc.* **2007**, *2* (8), 1896−1906.

(25) Chambers, M. C.; Maclean, B.; Burke, R.; Amodei, D.; Ruderman, D. L.; Neumann, S.; Gatto, L.; Fischer, B.; Pratt, B.; Egertson, J.; et al. A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* **2012**, *30* (10), 918−920.

(26) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999**, *20* (18), 3551−3567.



DOI: 10.1021/acs.jproteome.5b00579
*J. Proteome Res.* 2015, 14, 3568−3582

(27) Craig, R.; Beavis, R. C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **2004**, *20* (9), 1466−1467.
(28) Kim, S.; Pevzner, P. A. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat. Commun.* **2014**, *5*, 5277.
(29) Eng, J. K.; Jahan, T. A.; Hoopmann, M. R. Comet: an open-source MS/MS sequence database search tool. *Proteomics* **2013**, *13* (1), 22−24.
(30) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **2002**, *74* (20), 5383−5392.
(31) Shteynberg, D.; Deutsch, E. W.; Lam, H.; Eng, J. K.; Sun, Z.; Tasman, N.; Mendoza, L.; Moritz, R. L.; Aebersold, R.; Nesvizhskii, A. I. iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. *Mol. Cell. Proteomics* **2011**, *10* (12), M111.007690.
(32) Deutsch, E. W.; Mendoza, L.; Shteynberg, D.; Slagel, J.; Sun, Z.; Moritz, R. L. Trans-Proteomic Pipeline, a standardized data processing pipeline for large-scale reproducible proteomics informatics. *Proteomics: Clin. Appl.* **2015**, *9*, 745.
(33) Nesvizhskii, A. I.; Keller, A.; Kolker, E.; Aebersold, R. A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.* **2003**, *75* (17), 4646−4658.
(34) UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Res.* **2015**, *43* (Database issue), D204−D212.
(35) Lange, P. F.; Overall, C. M. TopFIND, a knowledgebase linking protein termini with function. *Nat. Methods* **2011**, *8* (9), 703−704.
(36) Lange, P. F.; Huesgen, P. F.; Overall, C. M. TopFIND 2.0–linking protein termini with proteolytic processing and modifications altering protein function. *Nucleic Acids Res.* **2012**, *40* (Database issue), D351−D361.
(37) Fortelny, N.; Yang, S.; Pavlidis, P.; Lange, P. F.; Overall, C. M. Proteome TopFIND 3.0 with TopFINDer and PathFINDer: database and analysis tools for the association of protein termini to pre- and post-translational events. *Nucleic Acids Res.* **2015**, *43* (Database issue), D290−D297.
(38) Vizcaíno, J. A.; Côté, R. G.; Csordas, A.; Dianes, J. A.; Fabregat, A.; Foster, J. M.; Griss, J.; Alpi, E.; Birim, M.; Contell, J.; et al. The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. *Nucleic Acids Res.* **2013**, *41* (Database issue), D1063−D1069.
(39) Boyle, E. I.; Weng, S.; Gollub, J.; Jin, H.; Botstein, D.; Cherry, J. M.; Sherlock, G. GO::TermFinder–open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics* **2004**, *20* (18), 3710−3715.
(40) Merico, D.; Isserlin, R.; Stueker, O.; Emili, A.; Bader, G. D. Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. *PLoS One* **2010**, *5* (11), e13984.
(41) Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N. S.; Wang, J. T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **2003**, *13* (11), 2498−2504.
(42) Maere, S.; Heymans, K.; Kuiper, M. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* **2005**, *21* (16), 3448−3449.
(43) Oesper, L.; Merico, D.; Isserlin, R.; Bader, G. D. WordCloud: a Cytoscape plugin to create a visual semantic summary of networks. *Source Code Biol. Med.* **2011**, *6*, 7.
(44) Lee, S.; Liu, B.; Lee, S.; Huang, S.-X.; Shen, B.; Qian, S.-B. Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109* (37), E2424−E2432.
(45) Fortelny, N.; Pavlidis, P.; Overall, C. M. The path of no return-Truncated protein N-termini and current ignorance of their genesis. *Proteomics* **2015**, *15* (14), 2547−2552.
(46) Yamakoshi, Y.; Hu, J. C.-C.; Iwata, T.; Kobayashi, K.; Fukae, M.; Simmer, J. P. Dentin sialophosphoprotein is processed by MMP-2 and MMP-20 in vitro and in vivo. *J. Biol. Chem.* **2006**, *281* (50), 38235−38243.
(47) Turk, V.; Stoka, V.; Vasiljeva, O.; Renko, M.; Sun, T.; Turk, B.; Turk, D. Cysteine cathepsins: from structure, function and regulation to new frontiers. *Biochim. Biophys. Acta, Proteins Proteomics* **2012**, *1824* (1), 68−88.
(48) Tsukamoto-Tanaka, H.; Ikegame, M.; Takagi, R.; Harada, H.; Ohshima, H. Histochemical and immunocytochemical study of hard tissue formation in dental pulp during the healing process in rat molars after tooth replantation. *Cell Tissue Res.* **2006**, *325* (2), 219−229.
(49) Sugiyama, Y.; Yamaguchi, M.; Kanekawa, M.; Yoshii, M.; Nozoe, T.; Nogimura, A.; Kasai, K. The level of cathepsin B in gingival crevicular fluid during human orthodontic tooth movement. *Eur. J. Orthod.* **2003**, *25* (1), 71−76.
(50) Romero-Quintana, J. G.; Frías-Castro, L. O.; Arámbula-Meraz, E.; Aguilar-Medina, M.; Dueñas-Arias, J. E.; Melchor-Soto, J. D.; Romero-Navarro, J. G.; Ramos-Payán, R. Identification of novel mutation in cathepsin C gene causing Papillon-Lefèvre Syndrome in Mexican patients. *BMC Med. Genet.* **2013**, *14*, 7.
(51) Repnik, U.; Starr, A. E.; Overall, C. M.; Turk, B. Cysteine Cathepsins Activate ELR Chemokines and Inactivate Non-ELR Chemokines. *J. Biol. Chem.* **2015**, *290* (22), 13800−13811.
(52) Butler, G. S.; Overall, C. M. Matrix metalloproteinase processing of signaling molecules to regulate inflammation. *Periodontol. 2000* **2013**, *63* (1), 123−148.
(53) Dufour, A.; Overall, C. M. Missing the target: matrix metalloproteinase antitargets in inflammation and cancer. *Trends Pharmacol. Sci.* **2013**, *34* (4), 233−242.
(54) Jágr, M.; Eckhardt, A.; Pataridis, S.; Mikšík, I. Comprehensive proteomic analysis of human dentin. *Eur. J. Oral Sci.* **2012**, *120* (4), 259−268.
(55) Farrah, T.; Deutsch, E. W.; Omenn, G. S.; Campbell, D. S.; Sun, Z.; Bletz, J. A.; Mallick, P.; Katz, J. E.; Malmström, J.; Ossola, R.; et al. A high-confidence human plasma proteome reference set with estimated concentrations in PeptideAtlas. *Mol. Cell. Proteomics* **2011**, *10* (9), M110.006353.
(56) Slany, A.; Meshcheryakova, A.; Beer, A.; Ankersmit, H. J.; Paulitschke, V.; Gerner, C. Plasticity of fibroblasts demonstrated by tissue-specific and function-related proteome profiling. *Clin. Proteomics* **2014**, *11* (1), 41.
(57) Hara, Y.; Kawasaki, N.; Hirano, K.; Hashimoto, Y.; Adachi, J.; Watanabe, S.; Tomonaga, T. Quantitative proteomic analysis of cultured skin fibroblast cells derived from patients with triglyceride deposit cardiomyovasculopathy. *Orphanet J. Rare Dis.* **2013**, *8*, 197.
(58) Hjorten, R.; Hansen, U.; Underwood, R. A.; Telfer, H. E.; Fernandes, R. J.; Krakow, D.; Sebald, E.; Wachsmann-Hogiu, S.; Bruckner, P.; Jacquet, R.; et al. Type XXVII collagen at the transition of cartilage to bone during skeletogenesis. *Bone* **2007**, *41* (4), 535−542.
(59) Suzuki, S.; Sreenath, T.; Haruyama, N.; Honeycutt, C.; Terse, A.; Cho, A.; Kohler, T.; Müller, R.; Goldberg, M.; Kulkarni, A. B. Dentin sialoprotein and dentin phosphoprotein have distinct roles in dentin mineralization. *Matrix Biol.* **2009**, *28* (4), 221−229.
(60) Tsuchiya, S.; Simmer, J. P.; Hu, J. C.-C.; Richardson, A. S.; Yamakoshi, F.; Yamakoshi, Y. Astacin proteases cleave dentin sialophosphoprotein (Dspp) to generate dentin phosphoprotein (Dpp). *J. Bone Miner. Res.* **2011**, *26* (1), 220−228.
(61) Wilhelm, M.; Schlegl, J.; Hahne, H.; Moghaddas Gholami, A.; Lieberenz, M.; Savitski, M. M.; Ziegler, E.; Butzmann, L.; Gessulat, S.; Marx, H.; et al. Mass-spectrometry-based draft of the human proteome. *Nature* **2014**, *509* (7502), 582−587.
(62) Uhlén, M.; Fagerberg, L.; Hallström, B. M.; Lindskog, C.; Oksvold, P.; Mardinoglu, A.; Sivertsson, Å.; Kampf, C.; Sjöstedt, E.; Asplund, A.; et al. Proteomics. Tissue-based map of the human proteome. *Science* **2015**, *347* (6220), 1260419.
(63) Deutsch, E. W.; Sun, Z.; Campbell, D.; Kusebauch, U.; Chu, C. S.; Mendoza, L.; Shteynberg, D.; Omenn, G. S.; Moritz, R. L. State of the Human Proteome in 2014/2015 As Viewed through PeptideAtlas: Enhancing Accuracy and Coverage through the AtlasProphet. *J. Proteome Res.* **2015**, DOI: 10.1021/acs.jproteome.5b00500.


DOI: 10.1021/acs.jproteome.5b00579
*J. Proteome Res.* 2015, 14, 3568−3582

(64) Ezkurdia, I.; Vázquez, J.; Valencia, A.; Tress, M. Analyzing the first drafts of the human proteome. *J. Proteome Res.* **2014**, *13* (8), 3854−3855.
(65) Lee, R. J.; Cohen, N. A. Taste receptors in innate immunity. *Cell. Mol. Life Sci.* **2015**, *72* (2), 217−236.
(66) Feldmesser, E.; Olender, T.; Khen, M.; Yanai, I.; Ophir, R.; Lancet, D. Widespread ectopic expression of olfactory receptor genes. *BMC Genomics* **2006**, *7* (1), 121.
(67) Varshavsky, A. The N-end rule pathway of protein degradation. *Genes Cells* **1997**, *2* (1), 13−28.
(68) Baggerman, G.; Vierstraete, E.; De Loof, A.; Schoofs, L. Gel-based versus gel-free proteomics: a review. *Comb. Chem. High Throughput Screening* **2005**, *8* (8), 669−677.
(69) Morsczeck, C.; Petersen, J.; Völlner, F.; Driemel, O.; Reichert, T.; Beck, H. C. Proteomic analysis of osteogenic differentiation of dental follicle precursor cells. *Electrophoresis* **2009**, *30* (7), 1175−1184.
(70) Bellac, C. L.; Dufour, A.; Krisinger, M. J.; Loonchanta, A.; Starr, A. E.; Auf dem Keller, U.; Lange, P. F.; Goebeler, V.; Kappelhoff, R.; Butler, G. S.; et al. Macrophage matrix metalloproteinase-12 dampens inflammation and neutrophil influx in arthritis. *Cell Rep.* **2014**, *9* (2), 618−632.
(71) Van Damme, P.; Arnesen, T.; Gevaert, K. Protein alpha-N-acetylation studied by N-terminomics. *FEBS J.* **2011**, *278*, 3822−3834.

## NOTE ADDED AFTER ASAP PUBLICATION

This paper was published ASAP on August 14, 2015 with ref 71 omitted. The corrected version was reposted on August 20, 2015.



DOI: 10.1021/acs.jproteome.5b00579
*J. Proteome Res.* 2015, 14, 3568−3582