Title

Top-Down Atmospheric Ionization Mass Spectrometry Microscopy Combined With Proteogenomics

Permalink
https://escholarship.org/uc/item/32072890

Journal
Analytical Chemistry, 89(16)

ISSN
0003-2700

Authors
Hsu, Cheng-Chih
Baker, Michael W
Gaasterland, Terry
et al.

Publication Date
2017-08-15

DOI
10.1021/acs.analchem.7b01096

Peer reviewed

Top-Down Atmospheric Ionization Mass Spectrometry Microscopy Combined With Proteogenomics

Journal:	*Analytical Chemistry*
Manuscript ID	ac-2017-01096c.R1
Manuscript Type:	Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Hsu, Cheng-Chih; National Taiwan University, Department of Chemistry Baker, Michael; University of California San Diego, Division of Biological Sciences Gaasterland, Terry; University of California San Diego Scripps Institution of Oceanography; University of California San Diego, Institute for Genomic Medicine Meehan, Michael; University of California, San Diego, Skaggs School of Pharmacy and Pharmaceutical Sciences Macagno, Eduardo; University of California, San Diego, Division of Biological Sciences Dorrestein, Pieter; University of California, San Diego, Skaggs School of Pharmacy

Top-Down Atmospheric Ionization Mass Spectrometry Microscopy Combined With Proteogenomics

Cheng-Chih Hsu,[†] Michael W. Baker,[‡] Terry Gaasterland,[§+] Michael J. Meehan,[l] Eduardo R. Macagno,[‡]* Pieter C. Dorrestein[l]*

[†] Department of Chemistry, National Taiwan University, Taipei 10617, Taiwan, [‡] Division of Biological Sciences, [+] Scripps Institution of Oceanography, [§] Institute for Genomic Medicine, and [l] Collaborative Mass Spectrometry Innovation Center, Skaggs School of Pharmacy and Pharmaceutical Science, University of California, San Diego, La Jolla, CA 92093.

ABSTRACT: Mass spectrometry-based protein analysis has become an important methodology for proteogenomic mapping by providing evidence for the existence of proteins predicted at the genomic level. However, screening and identification of proteins directly on tissue samples, where histological information is preserved, remain challenging. Here we demonstrate that the ambient ionization source, nanospray desorption electrospray ionization (nanoDESI), interfaced with light microscopy allows for protein profiling directly on animal tissues at the microscopic scale. Peptide fragments for mass spectrometry analysis were obtained directly on ganglia of the medicinal leech (*Hirudo medicinalis*) without in-gel digestion. We found that a hypothetical protein, which is predicted by the leech genome, is highly expressed on the specialized neural cells that are uniquely found in adult sex segmental ganglia. Via this top-down analysis, a post-translational modification (PTM) of tyrosine sulfation to this neuropeptide was resolved. This three-in-one platform, including mass spectrometry, microscopy, and genome mining, provides an effective way for mappings of proteomes under the lens of a light microscope.

The advance of cost-effective genomic sequencing has led to many completely sequenced genomes that are available to the public.[1-5] The increasing knowledge of genomic scaffolds is essential to the identification of gene-encoded putative proteins, and the experimental observation of these predicted proteins in various organisms has been growing enormously in the past decades.[6-9] Furthermore, those confirmations of protein expression provide a complementary tool to improve the annotation of genes, e.g. verification of open reading frames (ORFs), of the sequenced genome.[10] Such proteogenomic approaches have been utilized to advance the annotation of genomic datasets or the identification of specific gene clusters in various organisms.[9-15]

In the context of such exploration-driven efforts, mass spectrometry plays a crucial role in proteomic analysis.[6-18] For example, de novo tandem mass spectrometry (MS^n) analysis was utilized to locate the biosynthetic gene clusters for ten ribosomal and nonribosomal peptide natural products from the well-characterized microorganism.[14] Among all available ionization methods, electrospray ionization (ESI) and matrix-assisted laser desorption ionization (MALDI) are most widely used for the mass spectrometric characterization of proteins.[19,20] Both MALDI and ESI are considered as “soft ionization”, which preserves the native sequence of polypeptides in their ionic gas phases.[19,20] In this regard, intact (top-down) or digested (bottom-up) ionized proteins can be introduced into a mass spectrometer for peptide mass fingerprinting or be subjected to tandem mass analysis for subsequent annotation of sequences.[21,22]

However, in most of the mass spectrometry-based proteomics studies, appropriately prepared samples are prerequisite and critical to the quality of analysis.[22] These preparation steps mainly include: (1) tissue or cell homogenization and extractions, (2) protein digestions using proteases such as trypsin, and (3) molecular isolation processes such as liquid chromatography (LC) or polyacrylamide gel electrophoresis (PAGE). Although many high-throughput methods are available, valuable personnel time and expensive laboratory costs are indispensable for preparation of protein samples. Furthermore, the histological heterogeneity and topography are lost at sub-tissue level in most of the proteomics studies.

In the past 20 years, imaging mass spectrometry (IMS) has provided an excellent capability for visualizing label-free biomolecules in a spatial manner.[23-26] MALDI is currently the most widely used ionization sources for IMS. MALDI IMS is particularly superior in providing accurate spatial distribution of singly charged proteins from biological sections.[26-28] However, effective measurements of the MALDI IMS are greatly dependent on the quality of sample preparations, e.g. the uniformity of matrix deposition or the requirement of specialized substrates.[29] This generally limits the usefulness of MALDI IMS to cryo-sectioned biological samples, in order to preserve morphological details at the cellular level. Alternatively, IMS using ambient ionization sources provide complementary methods that only require minimal sample preparation.[30] Ambient ionization, such as desorption electrospray ionization (DESI), allows ionization of small molecules, i.e. M.W. < 1000 Dalton (Da), directly on the tissue surface without matrix deposition.[31-33] However, *situ* characterizations of larger biomolecules on biological samples are rarely reported using ambient ionization.[34-36] Recently, Hsu et al. showed that by interfacing with light microscopy, an ambient ionization source, nanospray desorption electrospray ionization (nanoDESI), is able to reveal developmental patterns of proteins associated with central nervous systems (CNS).[37] More importantly, this microscopy-guided approach allows top-down protein MS analysis, such that MS fragments containing amino acid (a.a.) sequences were obtained directly on tissue sections without enzymatic digestion. Herein, we present this hybrid technique, interfacing ambient mass spectrometry with light microscopy, to demonstrate rapid tissue-specific proteomic screening and subsequent *in situ* top-down identification of peptides present in the CNS of the medicinal leech.

EXPERIMENTAL SECTION

Microscopy NanoDESI. The hybrid microscopy ambient ionization source was used for rapid molecular screening directly on isolated medicinal leech ganglia. Details of the instrument setup have been described elsewhere.[37,38] In short, an inverted fluorescence microscope (Nikon DIAPHOT 300) was interfaced with a nanospray desorption electrospray ionization (nanoDESI) source for a two-in-one microscopy-guided mass spectrometry analysis. Sample plates were held on a z-axis translation stage, which was itself bolted onto the x-y stage of the microscope. The stage was then manipulated to position the ganglia in the area of interests under the liquid junction formed by the two fused silica capillary tubes of the nanoDESI instrument. The capillary tubes were flame-pulled from original 150/50 μm (O.D./I.D.) stock to ~50 μm O.D. (only at junction terminals), and a voltage of 2.0 kV (positive mode) was applied to the primary capillary tube. Throughout the experiments, a syringe pump was used to deliver a continuous flow of solvent to the capillary junction, at a rate of ~1.0 μL/min. The solvent for all nanoDESI-MS experiments was 65/35 (vol/vol) acetonitrile/0.1% formic acid aqueous solution. The sample stage was raised until the sample surface slightly contacted with the liquid droplet. A continuous stream of mobilized analytes as well as carrier solvent was then aspirated into the secondary capillary tube to create the electrospray. The distance (~1-2 mm) between the terminal end of the secondary capillary tube and the MS inlet was optimized for MS readouts and to keep the solvent flow-rate at the level of ~1.0 μL/min. The dynamic droplet size at the junction of the two capillaries and solvent spreading spot on the ganglionic surface were controlled within 100 μm in these experimental conditions. For each sampling spot, the ion signal (MS1) was accumulated for 2 minutes. After each measurement, the sample was removed from the liquid droplet (by lowering down the stage) and was allowed at least 1-minute rinse of pure blank solvent to prevent carry-over.

General Mass Spectrometry Settings. Electrospray ions generated by the home-built microscope-nanoDESI source or with a TriVersa Nanomate (Advion Biosystems) nanoESI source were directly introduced into a 6.4 T Finnigan ion trap quadrupole Fourier transform ion cyclotron resonance (LTQ-FT-ICR) mass spectrometer (Thermo) for MS analysis. Detailed operating parameters of nanoDESI are described in the previous section. HPLC-purified compounds were dissolved in spray solvent of 50:50 methanol/water containing 0.1% formic acid, and underwent nanoelectrospray ionization on a TriVersa Nanomate using a back pressure of 0.35-0.5 psi and the spray voltage of 1.3-1.45 kV. The instrument was first tuned to *m/z* 816.3 (+15 charge) using 2.0 μM commercially purchased bovine cytochrome c (Sigma-Aldrich) 65/35 (vol/vol) acetonitrile/0.1% formic acid aqueous solution before measurements on medicinal leech ganglia. For MS1 protein profile scanning using nanoDESI, positive ion spectra were collected in selected ion monitoring mode at 825.0 *m/z* with a 100 *m/z* isolation window. FT-ICR-MS spectra with a 50,000 resolution (at *m/z* 400) were collected using 4000-ms maximum ion inject time. For the mass determination of sulfopeptides, FT-ICR-MS spectra were collected with a resolution of 500,000 (at *m/z* 400), which is sufficient to resolve the mass differences between the PTMs of sulfation and phosphorylation.

Top-down MS Analysis. To obtain the a.a. sequence tags of the proteins of interests, top-down MS2 and higher tandem mass analyses were performed utilizing a normalized collision energy of 35% and an activation Q of 0.200 for both LTQ (isolation Δ*m/z*=3) and FT-ICR (isolation Δ*m/z*=8) detections. Product ion intensity using nanoDESI source were accumulated from multiple sample points to increase the signal-to-noise ratio. For protein analysis the charge states and monoisotopic mass of the ion fragments were determined by isotopic patterns including isotope Δ*m/z* and peak intensity ratio of each isotope cluster. The raw tandem FT-MS spectra were deconvoluted by Xtract (Thermo) into neutral monoisotopic spectra.[39] We also manually inspected and deconvoluted the raw spectra;[40] ion fragments of isotopic patterns that were not identified by Xtract were manually pulled out. The product ions of the same parent molecules were listed according to their monoistopic masses to search for mass shift matching to a single or multiple a.a. losses. Mass shift-based sequences containing five (or more) consecutive a.a. tags were then used to search for candidate proteins in the databases. Full a.a. sequence tags of the candidate proteins were furthered validated by cross-comparison with entire fragmenting products on ProSight PTM online platform.[41] Parent *m/z* values were verified with the theoretical *m/z* values of the candidates. Product ions of proteins were verified on the basis of the annotations to the monoisotopic *y* and *b* fragments. Isotopic distribution of predicted proteins and peptides fragments were calculated using IsoPro program basing on Yergey algorithm.[42] Prediction of sulfotyrosine sites was utilized by using SulfoSite (http://sulfosite.mbc.nctu.edu.tw/index.php).[43]

Proteome and Genome Mining. Protein BLAST engine at National Center for Biotechnology Information (NCBI) was used to search for candidates that contain the sequence tags suggested by the top-down MS analysis using parameters appropriate for short peptide queries (ungapped alignments and high e-values).[44] Sequences not found in known databases, or whose output candidates could not be validated by MS results, were compared to translated mRNA sequences predicted from a draft assembly for *Hirudo medicinalis*. Draft genome sequences were assembled from paired short reads using Velvet and PHRAP/CONSED[45,46] and given to GlimmerHMM to predict mRNAs.[47] Sequence tags were compared to mRNAs and genome sequences locally with BLAST's tblastn algorithm.

General HPLC Conditions and Purifications. All HPLC purifications were performed on an Agilent infinity 1260 HPLC equipped with a diode array detector, a manual injector, and a Biorad Model 2110 fraction collector. An analytical column (Aeris 3.6 µm WIDEPORE XB-C18 200 Å, 250 x 4.6 mm, Phenomenex) was used for sample analysis and purification at 25 °C. For the mobile phase, gradients of solvent A (98.0% H_2O, 1.9% acetonitrile, and 0.1% TFA) and solvent B (98.0% acetonitrile, 1.9% H_2O, and 0.1% TFA) were used with flow-rate of 1 mL/min. Extracts from both sex and somatic ganglia homogenates were fractioned by HPLC with a gradient of 5 to 50% solvent B in 36 minutes followed by 50 to 95% in 4 minutes.

Chemical Materials. Solvents for all experiments in this study, including acetonitrile, methanol, and water were purchased from Sigma-Aldrich (all HPLC grade). Acid adducts including formic acid and trifluoroacetic acid (TFA) were purchased from Sigma-Aldrich. Authentic standards, including serotonine, spermine, and spermidine were purchased from Sigma-Aldrich. Synthetic peptides, including leech sex-peptide (**2a**, product #371156, >98%) and its sulfation form 13-Tyr-SO_3H (#371245, >95%), were ordered and purchased from American Peptide Company, Inc.

Preparation of Leech Sample. Leeches used in these experiments were obtained from a *Hirudo medicinalis* breeding colony maintained in the Macagno laboratory. For the nanoDESI mass spectrometric analysis, segmental ganglia in body segments 4-7 were surgically excised from anesthetized adult leeches and pinned on 35 mm Sylgard-covered plastic dishes and the ganglia were desheathed of their connective-tissue capsule and the enveloping glial cells. Samples were kept at -20 °C for ~30 minutes while delivered to the mass spectrometry laboratory. Prior to the mass spectrometry measurements, each ganglion was rapidly rinsed 2 times with methanol (20 µL, ~10 seconds) and placed in an air-flow hood for 5 minutes to evaporate the methanol.

In Situ Hybridization. To further validate the distributions of the peptide, we assayed the histological distribution of sex peptide (**2a**) mRNA by *in situ* hybridization on whole embryos and adult de-sheathed ganglia according to previously described methods.[48,49] Briefly, single-stranded sense and antisense RNA probes containing digoxygenin (DIG)-labeled uridine were synthesized by *in vitro* transcription using linearized cDNA clones as a template. These probes were hybridized to fixed dissected embryos and adult ganglia. Unbound probe was removed by treatment with RNase, and bound probe was detected using alkaline phosphatase (AP)-conjugated antibodies to DIG.

RESULTS AND DISCUSSION

The medicinal leech, *Hirudo medicinalis*, is an important model organism for the study of nervous system function and development, and is also a unique species presently approved for use in human medical procedures.[50,51] Due to its relative accessibility and simplicity, the CNS of the medicinal leech has been extensively studied. The CNS of the leech consists of thirty-two bilateral neuromeres, of which the 4 anterior-most fuse to form the sub-esophageal ganglion and the 7 posterior-most fuse to form the caudal ganglion. The other individual ganglia in mid-body segments (SG1-SG21) are comprised of single bilateral neuromeres connected to each other by a bilateral pair of nerves and a single small medial nerve (Figure 1a). These ganglia contain about 400 neurons, except the two “sex” ganglia (SG5 and SG6) innervating the segments bearing the genitalia, which have about 350 additional neurosecretory cells. These cells gradually appear during post-embryonic development and are distributed throughout the sex ganglia exclusively. How activity of neuronal circuits that involves these complements of neurosecretory cells contributes to specific reproductive behaviors is currently unexplored.

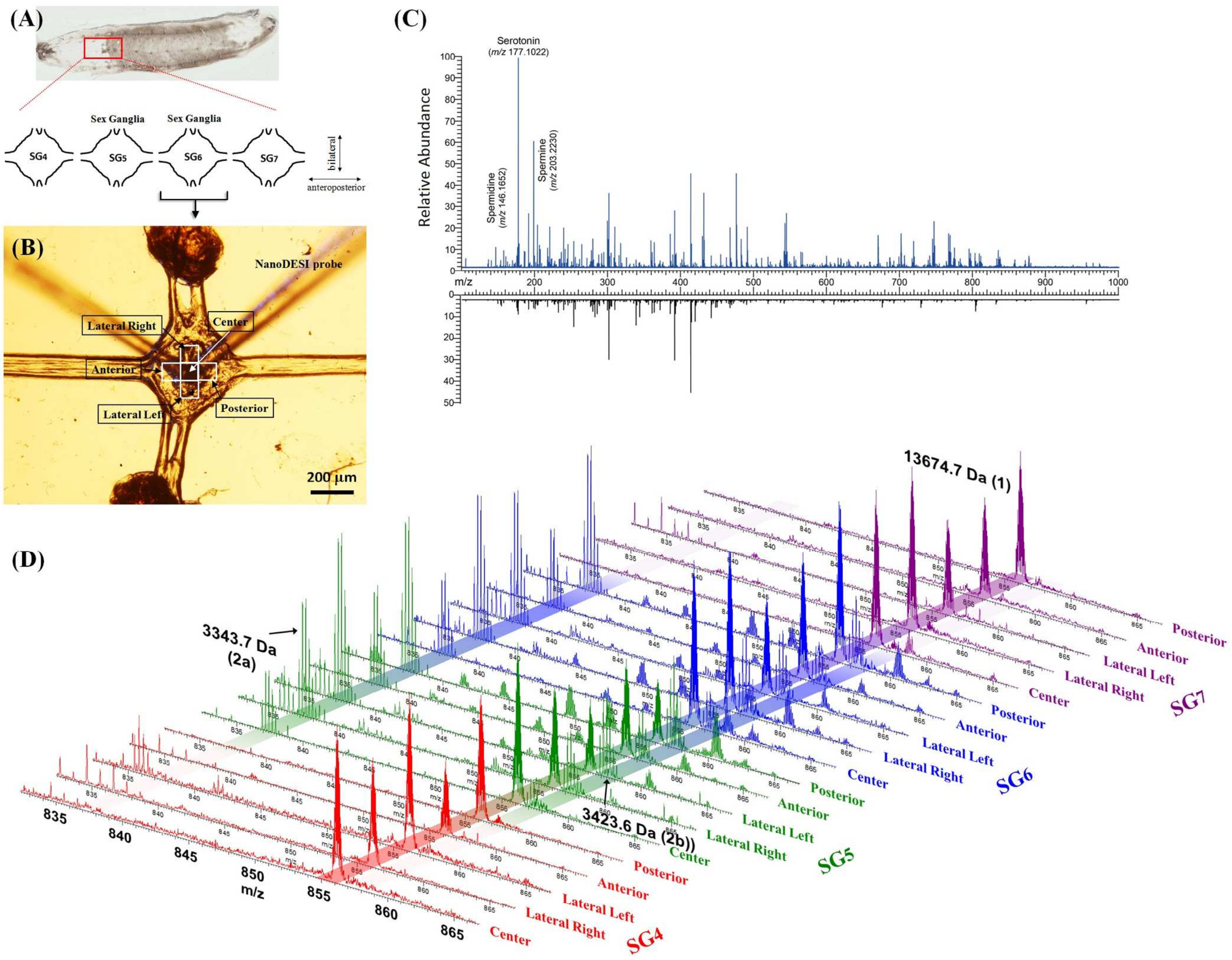


Figure 1. Microscopy ambient ionization mass spectrometry analysis on ganglia of the medicinal leech. (a) *Top:* photograph of a medicinal leech. The red rectangular region represents the four ganglia isolated for mass spectrometric analysis. *Bottom:* diagrams of SG4-SG7. SG5 and SG6 are "sex ganglia", which control the sex organs of the segments. (b) The microscopic snapshot of a SG6 interrogated by nanoDESI at the center of ganglia on the top surface. The white frames indicate anatomical positions at which mass spectra in (c) and (d) were collected. (c) *Top:* the mass spectrum acquired at the "center" position of SG6. *Bottom:* the background signal acquired on blank substrate gel. MS/MS of the annotated molecules are shown in supporting information Figure S1. (d) High resolution FT-MS spectra at the selected range (~ 830-870 *m/z*) taken at the anatomical positions specified in the micrograph (b). The MS spatial profiles reveal unique expressions of proteins neurohemorythin (13674.7 Da, compound **1**), leech sex peptide **2a** (3343.7 Da), and leech sex peptide **2b** (3423.6 Da) on the sex ganglia of medicinal leech.

The two sex segmental ganglia (SG5 and SG6) and two neighboring somatic ganglia (SG4 and SG7) were collected from live specimens and quickly frozen in their intact states for the subsequent nanoDESI-MS measurement. Five individual sampling spots for MS measurement were determined by microscopic imaging of each ganglion (Figure 1b). In all four ganglia, significant levels of 5-hydroxyltryptamine (serotonin, 177.1022 *m/z* [H^+]) were detected at most of the data points (Fig. 1c). This was expected, as several serotonergic cells, e.g., the very large Retzius neurons, are found in all leech segmental ganglia.[52] In addition to this neurotransmitter, polyamides such as spermidine and spermine were found throughout the samples. These have been known to be present ubiquitously in eukaryotes and in nervous systems.[53] Direct measurement of these low molecular weight organic polycations on mouse uterine sections during pregnancy was also reported using a nanoDESI source.[54]

In the previous study, it was shown that the nanoDESI source allows a gentle and highly-sensitive ionization of small-to-mid-sized proteins on tissue sections, with MS features that resemble conventional ESI sources.[37,55] Tissue-derived proteins as large as hemoglobin subunits (~14 kDa) were protonated (up to 21 H^+) and resolved by high-resolution mass spectrometry along with other molecular species of a lower mass range. Similarly, in this study, we sought to characterize larger molecules, with multiple charges at a higher *m/z* region. As shown in Figure 1d and supporting information Table S1, we found dozens of isotopic clusters of M.W. > 2 kDa, heterogeneously distributed across each ganglion. The deconvoluted monoisotopic masses of these compounds suggest that they are potential polypeptides associated with leech nervous systems. Among them, two sets of isotopic clusters, whose monoisotopic masses were 13674.74 Da (**1**), 3343.673 Da (**2a**), 3423.631 Da (**2b**), show distinct segmental specificities. Compound **1**, later identified as neurohemerythrin, was expressed ubiquitously, in all four ganglia at the same level. By contrast, compounds **2a** and **2b** were only detected in the sex ganglia (SG5 and SG6). To confirm this unique protein expression, we dissected twenty individual adult leech segmental ganglia and subjected them to *in situ* nanoDESI-MS and offline high-performance liquid chromatography (HPLC)-MS analysis. In both types of analysis, we did not detect any **2a** and **2b** ions in all of the "non-sex" ganglia (SG4 and SG7) while consistently found them in SG5 and SG6.

NanoDESI-MS provides not only an efficient way for rapid screenings of MS1 profiles, but also a capability to obtain top-down MS/MS that gives consecutive sequence tags of the leech peptides. The sequence tags can then be used to query the recently sequenced genome of *Hirudo verbana*, a close relative of *Hirudo medicinalis*, to identify the corresponding genes. To test the validity of this approach, we demonstrated how it works on the identification of neurohemerythrin (**1**). The ions at *m*/*z* 856.180 (*z*=16) acquired on SG4 to SG7 (Figure 1d) by nanoDESI were subjected to tandem MS using collision-induced dissociation (CID) and resulted in various fragments at *m*/*z* 883.444, 890.983, 899.587 (*z*=15) and 930.326 (*z*=14) corresponding to monoisotopic masses 13228.53, 13341.62, 13470.67, 13002.44 Da, respectively (Figure 2A). These masses together with the parent mass suggested multiple permutations of sequence tags containing [Gly+Phe]-Glu-Leu (or Ile)-[Pro+Glu]- (the order of tags in the brackets are not determined). This gave 8 possible sets of consecutive a.a. sequence tags that served as query units for search against the *Hirudo* genome. Search with NCBI protein BLAST engine found a prominent match (^{1}Gly-Phe-Glu-Ile-Pro-^{5}Glu-) to a leech protein neurohemerythrin known to be highly expressed in *H. medicinalis CNS*.[56] The intact protein mass rendered a negligible discrepancy of less than 1 ppm shift to the theoretical mass of neurohemerythrin. Furthermore, in the complementary HPLC-nanoESI-MS/MS results, another set of consecutive tags of neurohemerythrin (-^{82}Ala-Ser-Leu-Gly-Gly-Leu-Ser-^{88}Ser-) was identified (Figure 2A). The abovementioned information combined with observation of many other peptide fragment ions generated during CID of the neurohemerythrin (Figure S2) demonstrate unambiguously that the approach successfully identified this gene-encoded polypeptide.

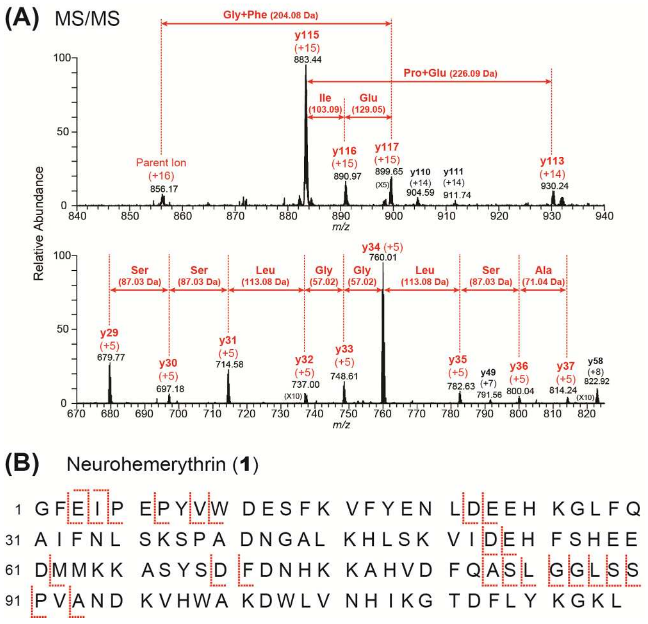


Figure 2. Top-down MS/MS analysis of neurohemerythrin. (A) Tandem mass spectra of neurohemerythrin (*m*/*z* 856.180) obtained by nanoDESI on the top of SG4 and SG7 (top) and from the nanoESI source (bottom). Consecutive sequence tags used for genome searching are shown at selected *m*/*z* regions. (B) The amino acid sequence and the fragmentation map of neurohemerythrin originated from MS/MS analysis of the protein at *z* = +16 to +19. Full MS/MS spectra and annotations of *y* and *b* ions of neurohemerythrin are exhibited in the supporting information Figure S2.

We then used this genome-assisted approach to characterize and identify the two sex ganglia-specific peptides **2a** at *m*/*z* 837.176 (*z*=4) and and **2b** at 857.166 (*z*=4). These molecules were separated by 79.958 Da implying a PTM of tyrosine sulfation was added in **2b**.[57] Notably, there is only a slight difference between sulfation- and phosphorylation-derived mass shifts, e.g. 79.9568 Da vs. 79.9663 Da (~3 ppm *m*/*z* difference for a +4 charge ion at *m*/*z* 857). Such a difference is typically considered as isobaric when using a low/medium resolution mass spectrometer.[58,59] However, such a mass discrepancy is in the range of significance when using a high resolution mass analyzer such as the orbitrap or the FTICR, as in this study with a resolving power as high as 500k (sufficient to resolve *m*/*z* difference of 2 ppm).[59] We thus carried out a multiple-point calibration to measure the exact masses of **2a** and **2b**. The observed *m*/*z* readouts of **2b** and its daughter ions agree well with the masses of the sulfated forms of **2a** (Figure 3 and S3). Moreover, the isotopic profiles of the **2b** ion clusters also present elevated abundance ratios at 3rd to 6th isotopomers, as a consequence of the contribution from natural ^{34}S isotope of the sulfate group (Table S2). These experimental results provided a conclusive piece of evidence favoring a considerable level of the sex peptide being displaced to its sulfated PTM form.

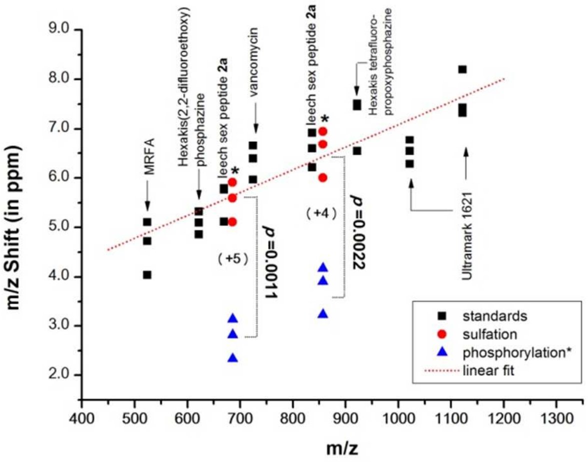


Figure 3. Determination of the exact mass of **2b** (sulfated form of **2a**). The calibration curve of the observed *m/z* readout shifts to the calculated *m/z* using standard compounds: MRFA (*m/z* 524.265), Hexakis(2,2-difluoroethoxy) phosphazine (*m/z* 622.029), synthetic leech sex peptide **2a** (*m/z* 669.742, z=5; *m/z* 836.931, z=4), vancomycin (*m/z* 724.7224, z=2), Hexakis tetrafluoropropoxy-phosphazine (922.010 *m/z*), Ultramark 1621 (1022.003 *m/z* and 1121.997 *m/z*). The values in the parentheses above are the calculated *m/z* of the denoted compounds. The corresponding *m/z* readout shifts of sex leech peptide **2b** to calculated sulfated (●) and phosphorylated (▲) PTM forms show statistical significance for both measurements in +4 and +5 charge states, indicating that the 80-Da PTM is contributed by a sulfation rather than phosphorylation.

To identify the peptides **2a** and **2b**, we first conducted a top-down MS/MS on SG5 and SG6 surfaces for ion *m/z* 837.176 of **2a** using nanoDESI MS. The resulting tandem mass spectrum provided profound information of ion fragments, rendering sequence tags of -Arg-Gln-Ser at the C-terminus (Figure S4). However, this annotation containing three tags was ineffective for genome search. Conventional experience has shown precursors of higher charge state are prone to give more informative ion products, as each peptide fragment has higher probability to possess charges so as to be measured by mass spectrometers.[60] High charge state precursor is also essential for a higher level tandem mass analysis (MS^n, n>2). Thus, we collected homogenate of 10 pairs of adult sex ganglia (SG5 and SG6) to extract the sex peptides for subsequent nanoESI MS measurements. The isolated sex peptides gave excellent CID peptidic fragment (*y* and *b* ions) as shown in Figure 4 (z=5) and supporting information Figure S5 (z=6). The MS2 of the non-sulfated peptide explicitly presented a series of sequence tags starting from the N-terminal denoting as Asn-Val-Ala-Leu (or Ile)-Arg-[388.18 Da]-Leu (or Ile)-Asp-[172.05 Da]. To obtain a more accurate series, we further investigated the fragments at MS^3 (by FTICR-MS) and MS^4 (by LTQ ion trap) levels. The MS^3 from *m/z* 558.453 (parent ion, z=6) → *m/z* 627.319 (+5) fragment, later known as the y_{25} ion of sex peptide **2a**, suggested that the gap of 172.05 Da is composed of -^{11}Gly-^{12}Asp- by the discovery of y_{16} ion (Figure S6). We further interrogated the peptide **2a** at MS^4 level on the ion of *m/z* 558.453 (parent ion, z=6) → *m/z* 627.319 (+5) → *m/z* 729.404 (+1), the fragment that contained both y_{25} and b_8 cleavages. The resulting spectrum ascertained that the 388-Da gap between y_{22} to y_{19} derived from -^{6}Pro-^{7}Tyr-^{8}Gln- (Figure S7). These findings are summarized as sequence tags of ^{1}Asn-Val-Ala-Leu (or Ile)-Arg-Pro-Tyr-Gln-Leu (or Ile)-Asp-Gly-^{11}Asp-.

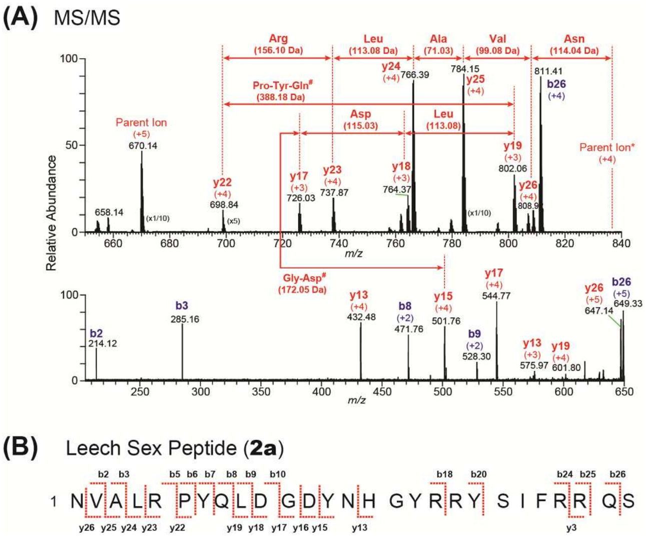
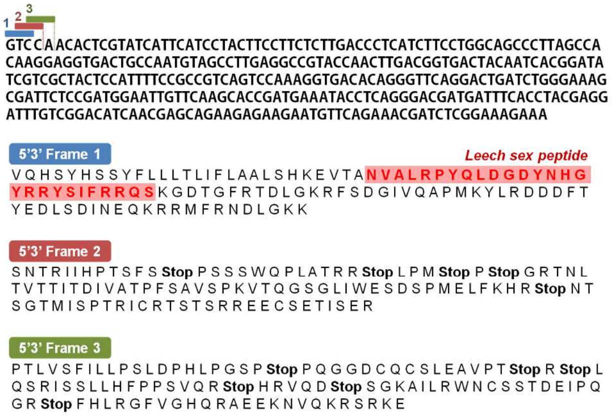


Figure 4. Top-down MS/MS analysis of leech sex peptide (**2a**). (A) Tandem mass spectra of leech sex peptide (**2a**) of ion *m/z* 669.943 (z=5) revealing 12 successive sequence tags from the N-terminal. The sequences tags were then utilized to blast against predicted *H. medicinalis* mRNA dataset. (B) The amino acid sequence and the fragmentation map of the hypothetical leech peptide originated from top-town tandem mass analyses (up to MS^4) of the peptide at z = +4 to +6.

These tags enable a successful match with a hypothetical protein sequence predicted by the *H. medicinalis* mRNA assembly (Figure 5). Draft genome sequences were subjected to Glimmer to identify putative mRNA sequences; their a.a. sequence translations were used for the search here.



GTCCAACACTCGTATCATTCATCCTACTTCCTTCTCTTGACCCTCATCTTCCTGGCAGCCCTTAGCCA
CAAGGAGGTGACTGCCAATGTAGCCTTGAGGCCGTACCAACTTGACGGTGACTACAATCACGGATA
TCGTCGCTACTCCATTTTCCGCCGTCAGTCCAAAGGTGACACAGGGTTCAGGACTGATCTGGGAAAG
CGATTCTCCGATGGAATTGTTCAAGCACCGATGAAATACCTCAGGGACGATGATTTCACCTACGAGG
ATTTGTCGGACATCAACGAGCAGAAGAGAAGAATGTTCAGAAACGATCTCGGAAAGAAA

5'3' Frame 1

Leech sex peptide

V Q H S Y H S S Y F L L L T L I F L A A L S H K E V T A **N V A L R P Y Q L D G D Y N H G**
Y R R Y S I F R R Q S K G D T G F R T D L G K R F S D G I V Q A P M K Y L R D D D F T
Y E D L S D I N E Q K R R M F R N D L G K K

5'3' Frame 2

S N T R I I H P T S F S **Stop** P S S S W Q P L A T R R **Stop** L P M **Stop** P **Stop** G R T N L
T V T T I T D I V A T P F S A V S P K V T Q G S G L I W E S D S P M E L F K H R **Stop** N T
S G T M I S P T R I C R T S T S R R E E C S E T I S E R

5'3' Frame 3

P T L V S F I L L P S L D P H L P G S P **Stop** P Q G G D C Q C S L E A V P T **Stop** R **Stop** L
Q S R I S S L L H F P P S V Q R **Stop** H R V Q D **Stop** S G K A I L R W N C S S T D E I P Q
G R **Stop** F H L R G F V G H Q R A E E K N V Q K R S R K E

Figure 5. Three-frame translation of the mRNA encoding the sex peptide **2a**. Excerpted nucleotide sequence and its forward strand reading frames (top) with the corresponding translated a.a. sequences (below). Sex peptide **2a,** highlighted in red in the verified sequence. The complete nucleotide and amino acid sequence of the sex peptide **2a** has been registered with GenBank, accession number KY707892.

Figure 6 shows the MS/MS annotations of **2b**. The fragmenting patterns of **2b** was similar to **2a**, the featuring Asn-Val-Ala-Leu-Arg-[388.18 Da]-Leu-Asp-[172.05 Da] sequence tag at the N-terminal was also found. Notably, a considerable amount of de-sulfated fragments were observed in the MS/MS of **2b**, indicating that the sulfate modification is labile to CID. This result was not surprising since the neutral loss of SO_3 (-80 Da) is very common in sulfopeptides.[61] There were four tyrosine residues in **2b**: Tyr-7, Tyr-13, Tyr-17, and Tyr-20. In the top-down tandem MS results of **2b**, both sulfated and de-sulfated y_{17} (see Figure 6) and y_{15} (see Figure S8) were found, whereas only de-sulfated form was measured as y_{13} ion. This suggested that the sulfation of **2b** occured on Tyr-13 (Figure 6B). To verify this annotation, it was submitted to SulfoSite, which computationally predicts sulfotyrisine sites within given peptide sequences basing on the accessible surface area surrounding sulfation residues.[43] SulfoSite predicted that Tyr-13 was the most probably sulfotyrisine in **2b**, with the highest Support Vector Machine (SVM) score among all tyrosine residues (see Supporting Information Table S3 and Figure S9). The SulfoSite prediction agreed with the tandem MS result, suggesting that a sulfate group was attached to Tyr-13 in **2b**.

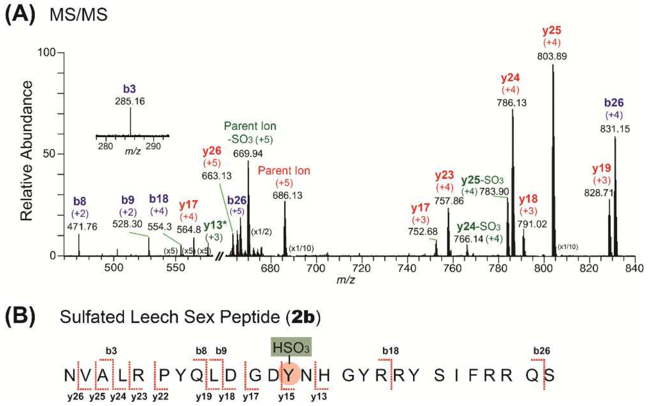


Figure 6. Top-down MS/MS analysis of leech sex peptide (**2b**) containing sulfotyrosine. (A) Tandem mass spectra of sulfated leech sex peptide (**2b**) of ion *m/z* 685.933 (z=5). The y ions (including y_{13}*) labeled in green are de-sulfated forms. Discovery of sulfated y_{17} and de-sulfated y_{13} implying a sulfotyrosine site at Tyr-13 (highlighted in red in (B)). (B) The fragmentation map of sulfopeptide **2b**. MS3 spectrum (parent ion → y_{25} ion→) is presented in the supporting information Figure S8.

To assay the histological distribution of the mRNA for sex peptide **2a**, we performed *in situ* hybridization on the adult and embryonic ganglia (Figure 7). Probe hybridization in non-sex ganglion 4 and 7 was found to be restricted to a few small cells scattered across the surface of the ganglion (Figure 7A and 7D). In contrast, the probe was found to hybridize with 100’s of small cells scattered across the ventral and dorsal surfaces of the sex ganglia (Figure 7B and 7C). Embryos on the other hand revealed no obvious differences in **2a** expression between sex and somatic ganglia (Figure 7E and 7F). These **2a**-positive cells have previously been identified as belonging to a population of peptidergic neurons found uniquely in ganglion 5 and 6 and which appear only late in embryonic development.[62] Named peripherally induced central (PIC) neurons, they are known to be induced by a peripheral signal associated with the developing male genitalia.[63]

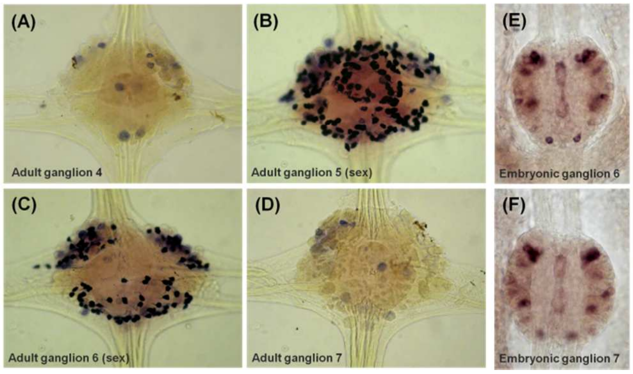


Figure 7. In situ hybridization staining of sex peptide **2a** on adult (A-D) and embryonic leech ganglia (E and F). Adult somatic ganglia: (A) and (D); adult sex ganglia: (B) and (C); embryonic sex and somatic ganglia: (E) and (F).

CONCLUSION

We have combined a proteogenomic strategy with ambient ionization mass spectrometry, microscopy, and genome mining. This combination supports and enables MS-based proteome profiling directly on the top of sample surface with only minimal sample pretreatment. The spatially mapped protein profile allows a site-specific top-down analysis to obtain the protein identities. Protein sequence tags are obtained by both mass spectrometric and genomic interrogation. Using medicinal leech as the model, we discovered a hypothetical peptide (**2a**), predicted from a draft leech genome and highly expressed in the adult leech sex ganglia. The result of immuno-staining suggests that this leech sex peptide **2a** is specific to the specialized neural cells unique to the adult sex segmental ganglia. Furthermore, the top-down analysis using high resolution mass spectrometry indicates that the sex peptide is susceptible to a PTM of sulfation on one of the tyrosine residues (Tyr-13). Thus, this microscopy-guided mass spectrometry-based proteogenomic platforms proves to be a powerful tool to exploratory biological sciences..

ASSOCIATED CONTENT

Supporting Information

The supporting information is available free of charge via the Internet at http://pubs.acs.org. The mass spectrometry data is available through http://gnps.ucsd.edu accession numbers MSV000080677. The complete nucleotide and amino acid sequence of the sex peptide **2a** has been registered with GenBank, accession number KY707892.

AUTHOR INFORMATION

Corresponding Author

E.R.M. Email: emacagno@ucsd.edu
P.C.D. Email: pdorrestein@ucsd.edu

ORCID

Pieter C Dorrestein: 0000-0002-3003-1030

Cheng-Chih Hsu: 0000-0002-2892-5326

Author Contributions

H.C.C. and B.M.W. performed the experiments. H.C.C. and G.T. analyzed the data. The manuscript was written with contributions from all authors. All authors have given approval to the final version of the manuscript.

Notes

The authors declare no competing financial interests.

ACKNOWLEDGMENT

This work was supported by NIH grant GM094802, AI095125, GM097509 and S10RR029121 (to the laboratory of P.C.D.), and NSF grant IOS-1355045 (to the laboratory of E.R.M). T.G. was supported by NIH grants ES024105 and EY022306. H.C.C. was supported by Ministry of Science and Technology (MOST), R.O.C. (Grant #: MOST 105-2113-M-002-004-MY2). We gratefully acknowledge Dr. Wen-Chi Chang in National Cheng Kung University for SulfoSite analysis.

REFERENCES

1. Shendure, J.; Ji, H. *Nat. Biotechnol*. **2008**, *26*, 1146-1153.
2. Mutz, K.-O.; Heilkenbrinker, A.; Lönne, M.; Walter, J.-G.; Stahl, F. *Curr. Opin. Biotechnol*. **2013**, *24*, 20-33.
3. Eid, J.; Fehr, A.; Gray, J.; Luong, K,; Lyle, J.; Otto, G.; Peluso, P.; Rank, D.; Baybayan, P.; Bettman, B.; et al. *Science* **2009**, *323*, 133-138.
4. Lander, E. S.; Linton, L. M.; Birren, B.; Nusbaum, C.; Zody, M. C.; Baldwin, J.; Devon, K.; Dewar, K.; Doyle, M.; FitzHugh, W.; et al. *Nature* **2001**, *409*, 860-921.
5. Adams M. D.; Celniker, S. E.; Holt, R. A.; Evans, C. A.; Gocayne, J. D.; Amanatides, P. G.; Scherer, S. E.; Li, P. W.; Hoskins, R. A.; Galle, R. F.; et al. *Science* **2000**, *287*, 2185-2195.
6. Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. *Electrophoresis* **1999**, *20*, 3551-3567.
7. Smith, J. C.; Northey, J. G.; Garg, J.; Pearlman, R. E.; Siu, K. W. *J. Proteome Res*. **2005**, *4*, 909-919.
8. Wang, X.; Zhang, B. *J. Proteome Res*. **2014**, 13, 2715-2723.
9. Gupta, N.; Tanner, S.; Jaitly, N.; Adkins, J. N.; Lipton, M.; Edwards, R.; Romine, M.; Osterman, A.; Bafna, V.; Smith, R. D.; et al. *Genome Res.* **2007**, *17*, 1362-1377.
10. Jaffe, J. D.; Berg, H. C.; Church, G. M. *Proteomics* **2004**, *4*, 59-77.
11. Shevchenko, A.; Jensen, O. N.; Podtelejnikov, A. V.; Sagliocco, F.; Wilm, M.; Vorm, O.; Mortensen, P.; Shevchenko, A.; Boucherie, H.; Mann, M. *Proc. Natl. Acad. Sci. USA* **1996**, *93*, 14440-14445.
12. Liu, W.-T.; Kersten, R. D.; Yang, Y.-L.; Moore, B. S.; Dorrestein, P. C. *J. Am. Chem. Soc*. **2011**, *133*, 18010-18013.
13. Khatun, J.; Yu, Y.; Wrobel, J. A.; Risk, B. A.; Gunawardena, H. P.; Secrest, A.; Spitzer, W. J.; Xie, L. ; Wang, L.; Chen, X.; Giddings, M. C. *BMC Genomics* **2013**, *14*, 141.
14. Kersten, R. D.; Yang Y. L.; Xu, Y.; Cimermancic, P.; Nam, S. J.; Fenical, W.; Fischbach, M. A.; Moore, B. S.; Dorrestein, P. C. *Nat. Chem. Biol.* **2011**, *7*, 794-802.
15. Kim, M.-S.; Pinto, S. M.; Getnet, D.; Nirujogi, R. S.; Manda S. S.; Chaerkady, R.; Madugundu, A. K.; Kelkar, D. S.; Isserlin, R.; Jain, S.; et. al. *Nature* **2014**, *509*, 575-581.
16. Wilhelm, M.; Schlegl, J.; Hahne, H.; Gholami, A. M.; Lieberenz, M.; Savitski, M. M.; Ziegler, E.; Butzmann, L.; Gessulat S.; Marx, H.; et al. *Nature* **2014**, *509*, 582-587.
17. Skinner, O. S.; Havugimana, P. C.; Haverland, N. A.; Fornelli, L.; Early, B. P.; Greer, J. B.; Fellers, R. T.; Durbin, K. R.; Do Vale, L. H.; Melani, R. D.; et al. *Nat. Methods*, **2016**, *13*, 237-240.
18. Hoshino, A.; Costa-Silva, B.; Shen, T.-L.; Rodrigues, G.; Hashimoto, A. *Nature*. **2015**, *527*, 329-335.
19. Fenn, J. B.; Mann, M.; Meng, C. K.; Wong, S. F.; Whitehouse, C. M. *Science* **1989**, *246*, 64-71.
20. Hillenkamp, F.; Karas, M.; Beavis, R. C.; Chait, B. T. *Anal. Chem.* **1991**, *63*, 1193A-1203A.
21. Kelleher, N. L.; Lin, H. Y.; Valaskovic, G. A.; Aaserud, D. J.; Fridriksson, E. K.; McLafferty, F. W. *J. Am. Chem. Soc.* **1999**, *121*, 806–812.
22. Domon, B.; Aebersold, R. *Science*, **2006**, *312*, 212-217.
23. Amstalden van Hove, E. R.; Smith, D. F.; Heeren R. M. A. *J. Chromator A.* **2010**, *1217*, 3946-3954.
24. Watrous, J. D.; Alexandrov, T.; Dorrestein, P. C. *J. Mass Spectrom*. **2011**, *46*, 209-222.
25. Ganesana, M.; Lee, S. T.; Wang, Y.; Venton, B. J. *Anal. Chem.* **2017**, *89*, 314-341.
26. Schwamborn, K.; Caprioli, R. M. *Nat. Rev. Cancer* **2010**, *10*, 639-646.
27. Schöne, C.; Höfler, H.; Walch, A. *Clin. Biochem*. **2013**, *46*, 539-545.
28. Cornett, D. S.; Reyzer, M. L.; Chaurand, P.; Caprioli, R. M. *Nat. Methods* **2007**, *4*, 828-833.
29. Northen, T. R.; Yanes, O.; Northen, M. T.; Marrinucci, D.; Uritboonthai, W.; Apon, J.; Golledge, S. L.; Nordström, A.; Siuzdak, G. *Nature* **2007**, *449*, 1033-1036.
30. Hsu, C.-C.; Dorrestein, P. C. *Curr. Opin. Biotechnol.* **2015**, *31*, 24-34.
31. Takats, Z.; Wiseman, J. M.; Gologan, B.; Cooks, R. G. *Science* **2004**, *306*, 471-473.
32. Wiseman, J. M.; Ifa, D. R.; Song, Q.; Cooks R. G. *Angew. Chem. Int. Ed.* **2006**, *45*, 7188-7192.
33. Feider, C. L.; Elizondo, N.; Eberlin, L. S. *Anal. Chem.*, **2016**, *88*, 11533-11541.
34. Kiss, A.; Smith1 D. F.; Reschke B. R.; Powell, M. J.; Heeren, R. M. A. *Proteomics*, **2014**, *14*, 1283-1289.
35. Sarsby, J.; Martin, N. J.; Lalor, P. F.; Bunch, J.; Cooper, H. J. *J. Am. Soc. Mass Spectrom.* **2014**, 25, 1953-1961.
36. Randall, E. C.; Bunch, J.; Cooper, H. *J. Anal. Chem.* **2014**, *86*, 10504-10510.
37. Hsu, C.-C.; White, N. M.; Hayashi, M.; Lin, E.C.; Poon, T.; Banerjee, I.; Chen, J.; Pfaff, S. L.; Macagno, E. R.; Dorrestein, P. C. *Proc. Natl. Acad. Sci. U.S.A.* **2013**, *110*, 14855-14860.
38. Mascuch, S. J.; Moree, W. J.; Hsu, C.-C.; Turner, G. G.; Cheng, T. L.; Blehert, D. S.; Kilpatrick, A. M.; Frick, W. F.; Meehan, M. J.; Dorrestein, P. C.; et al. *PloS one* **2015**, *10*, e0119668.
39. Zabrouskov, V.; Senko, M. W.; Du, Y.; Leduc, R. D.; Kelleher, N. L. *J. Am. Soc. Mass Spectrom.* **2005**, *16*, 2027-2038.
40. Mann, M.; Meng, C. K.; Fenn, J. B. "Interpreting mass spectra of multiply charged ions." *Anal. Chem.* **1989**, *61*, 1702-1708.
41. LeDuc, R. D.; Taylor, G. K.; Kim, Y. B.; Januszyk, T. E.; Bynum, L. H.; Sola, J. V.; Garavelli, J. S.; Kelleher, N. L. *Nucleic Acids Res.* **2004**, *32*, W340-W345.
42. Yergey, J. A. *Int. J. Mass Spectrom.* **1983**, *52*, 337-349.
43. Chang, W.-C.; Lee, T.-Y.; Shien, D.-M.; Hsu, J.-B. K.; Horng, J.-T.; Hsu, P.-C.; Wang, T.-Y.; Huang H.-D.; Pan, R.-L. *J Comput. Chem.*, **2009**, *30*, 2526-2537.
44. Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. *J. Mol. Biol.* **1990**, *215*, 403–410.
45. Zerbino, D. R. *Curr. Protoc. Bioinformatics* **2010**, *31*, 11.5.1–11.5.12.
46. Gordon, D.; Green, P. *Bioinformatics* **2013**, *29*, 2936-2937.
47. Majoros, W. H.; Pertea, M.; Salzberg, S. L. *Bioinformatics* **2004**, *20*, 2878-2879.

48. Nardelli-Haefliger, D.; Shankland, M. *Development* **1992**, *116*, 697–710.

49. Dykes, I. M.; Freeman, F. M.; Bacon, J. P.; Davies, J. A. *J. Neurosci.* **2004**, *24*, 886-894.

50. Wells, M. D.; Manktelow, R. T.; Boyd, J. B.; Bowen, V. *Microsurgery* **1993**, *14*, 183–186.

51. Rados, C. *FDA Consum.* **2004**, *38*, 9.

52. Lent, C. M.; Zundel, D.; Freedman, E.; Groome, J. R. *J. Comparative Physiol. A*, **1991**, *168*, 191-200.

53. Laube, G.; Bernstein, H.-G.; Wolf, G.; Veh, R. W. *J. Comparative Neurology*, **2002**, *444*, 369-386.

54. Lanekoff, I.; Burnum-Johnson, K.; Thomas, M.; Short, J.; Carson, J. P.; Cha, J.; Dey, S. K.; Yang, P.; Prieto Conaway, M. C.; Laskin, J. *Anal. Chem*. **2013**, *85*, 9596-9603.

55. Hsu, C.-C.; Chou, P.-T.; Zare, R. N. *Anal. Chem.* **2015**, *87*, 11171-11175.

56. Vergote, D.; Sautiere, P.-E.; Vandenbulcke, F.; Vieau, D.; Mitta, G.; Macagno, E. R.; Salzet, M. *J. Biol. Chem.* **2004**, *279*, 43828-43837.

57. Kanan, Y.; Al-Ubaidi, M. R. *JSM Biotechnol. Bioeng*. **2013**, *1*, 1003.

58. Zhang, Y.; Jiang, H.; Go, E. P.; Desaire, H. *J. Am. Soc. Mass Spectrom.* **2006**, *17*, 1282-1288.

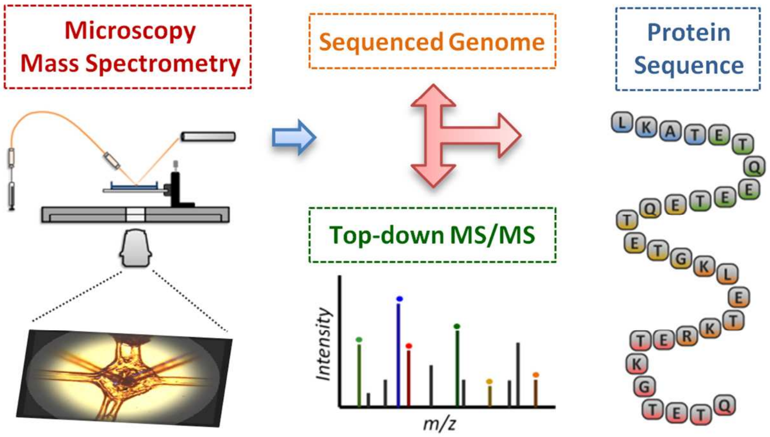
59. Bossio, R. E.; Marshall, A. G. *Anal. Chem*. **2002**, *74*, 1674-1679.

60. Garcia, B. A. *J. Am. Soc. Mass Spec.* **2010**, *21*, 193-202.

61. Monigatti F.; Hekking B.; Steen H. *Biochimica et Biophysica Acta, Proteins and Proteomics.* **2006**, *1764*, 1904–13.

62. Baptista, C. A.; Gershon, T. R.; Macagno, E. R. *Nature* **1990**, *346,* 855-858.

63. Becker, T. S.; Bothe, G.; Berliner, A. J.; Macagno, E. R. *Development* **1996**, *122*, 2331-2337.


Microscopy
Mass Spectrometry
Sequenced Genome
Top-down MS/MS
Intensity
m/z
Protein
Sequence
LKATETQEETEQTETGKLETKRETKGTETQ