ImShot: An Open-Source Software for Probabilistic Identification of Proteins *In Situ* and Visualization of Proteomics Data

Authors

Wasim Aftab, Shibojyoti Lahiri, and Axel Imhof

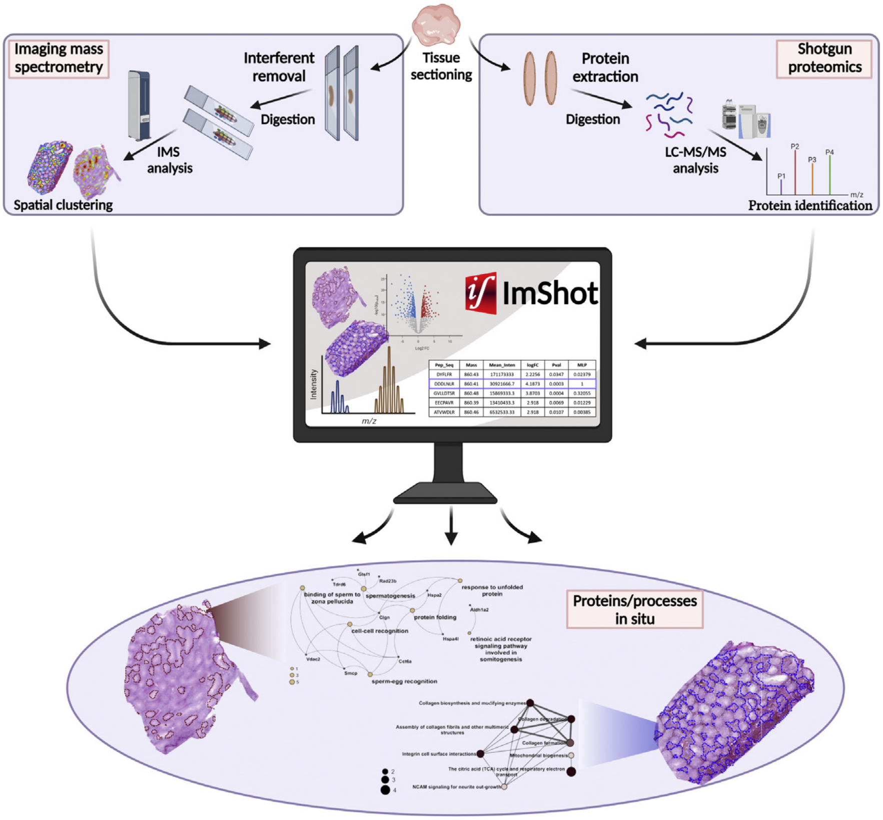
Correspondence

shibojyoti.lahiri@med.uni-muenchen.de; imhof@lmu.de

In Brief

ImShot is the first systematic software that integrates data from IMS and shotgun proteomics (LC–MS) to identify proteins *in situ*. The software is designed to identify proteins that are particularly responsible for establishing a diseased state. The software usage is independent of any mass spectrometry platform used to generate the data and can also be used for independent proteomics data analysis. ImShot is provided as a desktop application, thereby making its run free of external influencing factors.

Graphical Abstract


Imaging mass spectrometry
Interferent removal
Tissue sectioning
Digestion
IMS analysis
Spatial clustering
Protein extraction
Shotgun proteomics
Digestion
LC-MS/MS analysis
P1
P2
P3
P4
m/z
Protein identification
ImShot
Intensity
m/z
Proteins/processes in situ
Gtsf1
Tdrd6
Rad23b
binding of sperm to zona pellucida
spermatogenesis
Hspa2
response to unfolded protein
Aldh1a2
Clgn
protein folding
cell-cell recognition
Hspa4l
retinoic acid receptor signaling pathway involved in somitogenesis
Vdac2
Cct6a
Smcp
sperm-egg recognition
1
3
5
Collagen biosynthesis and modifying enzymes
Collagen degradation
Assembly of collagen fibrils and other multimeric structures
Collagen formation
Integrin cell surface interactions
Mitochondrial biogenesis
The citric acid (TCA) cycle and respiratory electron transport
NCAM signaling for neurite out-growth
2
3
4

Highlights

- ImShot is a systematic software integrating IMS and LC–MS.
- ImShot is designed for *in situ* identification of proteins involved in diseases.
- It is the first software performing the critical task of deisotoping IMS spectra.
- Modular structure of ImShot allows users to analyze LC–MS data independently.
- ImShot can be run by one-time download and plug and play.

2022, Mol Cell Proteomics *21*(*6*), 100242

© 2022 THE AUTHORS. Published by Elsevier Inc on behalf of American Society for Biochemistry and Molecular Biology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

https://doi.org/10.1016/j.mcpro.2022.100242

ImShot: An Open-Source Software for Probabilistic Identification of Proteins *In Situ* and Visualization of Proteomics Data

Wasim Aftab[1,2], Shibojyoti Lahiri[1,*], and Axel Imhof[1,*]

Imaging mass spectrometry (IMS) has developed into a powerful tool allowing label-free detection of numerous biomolecules *in situ*. In contrast to shotgun proteomics, proteins/peptides can be detected directly from biological tissues and correlated to its morphology leading to a gain of crucial clinical information. However, direct identification of the detected molecules is currently challenging for MALDI–IMS, thereby compelling researchers to use complementary techniques and resource intensive experimental setups. Despite these strategies, sufficient information could not be extracted because of lack of an optimum data combination strategy/software. Here, we introduce a new open-source software ImShot that aims at identifying peptides obtained in MALDI–IMS. This is achieved by combining information from IMS and shotgun proteomics (LC–MS) measurements of serial sections of the same tissue. The software takes advantage of a two-group comparison to determine the search space of IMS masses after deisotoping the corresponding spectra. Ambiguity in annotations of IMS peptides is eliminated by introduction of a novel scoring system that identifies the most likely parent protein of a detected peptide in the corresponding IMS dataset. Thanks to its modular structure, the software can also handle LC–MS data separately and display interactive enrichment plots and enriched Gene Ontology terms or cellular pathways. The software has been built as a desktop application with a conveniently designed graphic user interface to provide users with a seamless experience in data analysis. ImShot can run on all the three major desktop operating systems and is freely available under Massachusetts Institute of Technology license.

Proteomic studies over the years have aimed at understanding the functional landscape of cells by optimum mapping of protein profiles at steady state and following a variety of perturbations in space and time. In addition to conventional LC–MS, imaging mass spectrometry (IMS) provides a new dimension by enabling the detection of proteins *in situ*. Because of the enormous cellular heterogeneity inside tissues, a better understanding of the spatial distribution of proteins has become imperative for a clearer interpretation of human diseases (1–4). Indeed, a wide variety of studies show that effects of and/or responses to diseases could be cell type specific (5–7). Furthermore, diseases like cancer show very precise region-specific molecular alterations. In addition to distinct cell types and molecular profiles that characterize tumor, stroma, and the vasculature, intratumoral heterogeneity is often considered to be of prime importance in assessing the clinical status of malignant tumors (8–11). Therefore, identifying factors *in situ* that characterize the diseased state as compared with the healthy one is instrumental in improved assessment of diseases and formulation of better treatment strategies.

Since its inception, IMS studies have successfully mapped molecular profiles to tissue morphologies in the disease context (12–15). Discovery of biomarkers and biological categorization of relevant diseases was also possible through IMS-based investigations (16–21). However, most of these studies included complementary validation of IMS data since direct identification of molecules was not possible. Although metabolic profiles of tissues can now be identified in a relatively better way than before https://metaspace2020.eu/; (Accessed on 2021/05/17) (22), large-scale peptide identifications can still not be performed by IMS leading to suboptimal understanding of functional proteomic profiles of different cell types within tissues. Therefore, complementation of MALDI–IMS with orthogonal shotgun proteomics has been adopted as a feasible approach in the recent past (23–28).

The combination of these two orthogonal technologies has led from poor to substantial identification of proteins in a contextual manner. However, there is still a lack of appropriate strategies that could effectively combine data from these two platforms into an efficient screening module of proteins *in situ*. In fact, most of the attempts to combine the two modalities

From the [1]Biomedical Center, Protein Analysis Unit, Faculty of Medicine, Ludwig-Maximilians-Universität München, Planegg-Martinsried, Germany; [2]Graduate School for Quantitative Biosciences (QBM), Ludwig-Maximilians-Universität Munich, Munich, Germany
*For correspondence: Shibojyoti Lahiri, shibojyoti.lahiri@med.uni-muenchen.de; Axel Imhof, imhof@lmu.de.

© 2022 THE AUTHORS. Published by Elsevier Inc on behalf of American Society for Biochemistry and Molecular Biology.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

https://doi.org/10.1016/j.mcpro.2022.100242

involved considerable manual curation leading to very limited number of identified discriminative masses (23, 24, 29). More successful approaches were associated with measurements of *in situ* tryptic peptides with very high mass accuracy (comparable to LC–MS) leading to the analysis method being particularly resource intensive (25, 26, 30–32). None of the approaches developed so far has a defined integrated "one-in-all" workflow/software in the form of a graphic user interface (GUI) leading to the tedious task of combining multiple platforms with substantial manual input. Only recently, a command line–based package (32) has been proposed to identify IMS peptides from very high-resolution IMS data, but the workflow does not account for deisotoping IMS spectra, which as we demonstrate in this article could be pivotal in removing false-positive identifications. In addition, these approaches were exclusive of the two-group comparison scenario (healthy *versus* diseased), thereby providing very limited biological insights as part of a data analysis pipeline.

In this report, we introduce ImShot, a conveniently designed software that can be deployed as a screen for probabilistic identifications of proteins *in situ* in a disease *versus* healthy context. ImShot initially processes data from both IMS and LC–MS to filter for experimental, analytical, and isotopic contaminants. The individual mass lists thus created from the two complementary techniques are matched within a user-specified tolerance. Within the results, ambiguity arising from one-to-many mass annotations is largely reduced by ranking the peptide masses according to the novel scoring system that identifies the most likely protein *in situ*. In addition, there are options to validate the likelihood of peptide identification from IMS computationally. The software does this by using in-built functional validation tools like Gene Ontology (GO) and pathway analysis that associates biological processes (BPs) to the most likely region within a tissue specimen. ImShot has been developed using a modular structure that allows the user and/or the developer to customize their individual needs. As a result, it also allows a user to use this software for analyzing and visualizing LC–MS data separately. Finally, we have developed this whole package into an open source, convenient, and user-friendly desktop application using web technologies on Electron framework https://www.electronjs.org/; (Accessed on 2021/04/24) that operates on all the major operating systems (OSs).

EXPERIMENTAL PROCEDURES

Methods

ImShot is currently applicable on IMS datasets produced by measuring *in situ* generated peptides. As ImShot is primarily a data integration software, it can also be applied to IMS datasets of other types of molecules (*e.g.*, lipids, metabolites, etc) as long as there is an LC–MS counterpart to it. The example dataset used here to describe the features and mode of operation of the software are available *via* ProteomeXchange with identifier PRIDE: PXD022870 (7). The datasets were created to better understand the establishment of the complicated male infertility phenotype. Dysregulation of spermatogenesis involves a vast range of cell types that characterize various regions of the testicular tissue. The investigations sought to find proteins and molecular processes from these several tissue locations that could have resulted in the infertile phenotype. The experiments involved IMS and LC–MS/MS measurements on serial sections of healthy (WT) and transgenic (AROM+, infertile) mouse testes.

MALDI–IMS

WT and AROM+ tissues (fresh frozen) were cryosectioned into 12 μm thick sections and thaw-mounted on glass slides with indium-tin oxide coatings (Bruker Daltonik GmbH). A solution of 25 ng/μl trypsin in 20 mM ammonium bicarbonate was used to generate peptides *in situ* after thorough washing to remove interferents from tissues. Sections sprayed with the solution were incubated at 50 °C for 2 h and 30 min in a humid environment followed by matrix (10 mg/ml α-cyano-4-hydroxycinnamic acid in 70% acetonitrile and 1% trifluoroacetic acid) spray.

Imaging experiments were carried out in a rapifleX MALDI Tissuetyper MALDI-TOF/TOF mass spectrometer (Bruker Daltonik GmbH) equipped with a SmartBeam 3G laser. Positive reflector mode was used to measure peptide masses within a range of 600 to 3200 Da and with a spatial resolution of 25 μm. The measurements were externally calibrated using a commercial peptide calibrant combination (Peptide calibration standard II; Bruker Daltonik GmbH) that covered the aforementioned measured mass range and were spotted at several locations on the same target slide as the tissues. The generated data were segregated into different spatial clusters corresponding to defined tissue compartments. In this particular case, we have used SCiLS Lab (33) to generate the spatial peptide clusters and their corresponding distinguishing masses, but it is also possible to use any other software/algorithm to generate the aforesaid clusters and mass lists. However, to ensure identical experimental conditions and reduce the chances of ending up with overlapping clusters, it is recommended to have tissues from both the control and experimental/diseased condition on the same IMS slide.

Shotgun Proteomics (LC–MS)

A total of approximately 1 mg of mouse testis tissue was used to extract proteins and generate peptides using the iST Sample Preparation Kit (PreOmics) according to the manufacturer's protocol. The peptides were separated on a Thermo Fisher Scientific Ultimate 3000 nanoLC system and ionized in a nanoESI source and identified on-line with a QExactive HF mass spectrometer (Thermo Fisher Scientific).The mass spectrometer was operated in positive ionization mode using the TOP10 method, detecting eluting peptide ions between *m/z* 375 and 1600 and performing MS/MS analysis on up to 10 most intense precursor ions. Internally calibrated mass spectra were used to determine the masses of ambient siloxanes. Precursors were chosen on the basis of their intensity from all signals with a charge state ranging from 2+ to 5+, isolated in a *m/z* 2 window, and fragmented with a normalized collision energy of 27%. To avoid repeated fragmentation of the same peptide ion, dynamic exclusion was set to 20 s.

The raw data were initially processed for protein identification and quantification in MaxQuant software [Computational Systems Biochemistry (Cox Lab), Max Planck Institute of Biochemistry] (34). Currently, ImShot can read the output files (proteingroups.txt and peptides.txt) from MaxQuant processing only. Output from any other processing software can be used provided those are converted to a similar file format as mentioned before.

Computational Methods

Algorithm Overview—The ImShot software employs an algorithm that identifies peptides from IMS datasets based on comparison with corresponding proteomic data and peptide ranking according to the novel scoring method. It first filters IMS and LC–MS data to remove all false-positive entries arising because of different types of contaminants (for details, refer to the data processing sections of LC–MS and MALDI–IMS). Mass lists from IMS clusters are then matched with the proteomics dataset within a user-specified tolerance range. Because of the incompatibility of mass resolution between MALDI–IMS and LC–MS systems, the resulting one-to-many mass annotations are resolved by rating the peptide masses according to our new most likely peptide (MLP) scoring system. ImShot's built-in functional validation tools (*viz.*, GO and pathway analysis) can then be used to not only computationally assess the likelihood of the aforesaid identification from IMS but also for analyzing proteomics data separately. BPs can be associated with the most probable cell types/tissue compartments, thereby assessing the relevance of the algorithm. The modular structure of ImShot permits the user and/or the developer to flexibly adapt the software to their individual needs. This is reflected in the seamless usage of this software for analyzing and visualizing LC–MS data separately. To ensure smooth and hassle-free user experience, ImShot has been developed into a user-friendly desktop application that operates on all the major OSs.

Backend Computing—All the backend computing is done in R. The software has three main sections, *viz.*, data processing and statistical analysis, data integration, and functional assessment/validation. Each of these sections are subdivided into modules that individually carry out the desired tasks for the user (Fig. 1).

Data Processing (LC–MS)—For LC MS/MS data processing, ImShot reads either proteingroups.txt or peptides.txt (from MaxQuant run output) based on user input. In addition, to facilitate wider usage, the users can use other tab-delimited files that have columns with intensity-based absolute quantification/label-free quantitation/intensity headers for two groups and also a column named "Fasta headers." It then filters the identified protein list according to the following steps (Fig. 1):

i. In the first step, this module removes proteins classified as contaminants from the dataset. Within ImShot, the algorithm searches the columns "Reverse," "Potential contaminant," and "Only identified by site" for positive entries and removes all rows that contain the + sign. Specifically in case of peptides.txt file generated by using carbamidomethyl as "fixed modifications" in the MaxQuant run, the user can select "TRUE" from the drop-down menu in ImShot, which then adjusts for carbamidomethylation of cysteines (mass reduction of 57.02 Da). For datasets that do not have carbamidomethyl as "fixed modifications," the user can opt for the "FALSE" option.

ii. Following that, the algorithm takes care of any blank proteins that are included in the MaxQuant list of identified proteins. The program removes rows (proteins) from the dataset that contain only zeros for all replicates in all conditions. In addition, ImShot

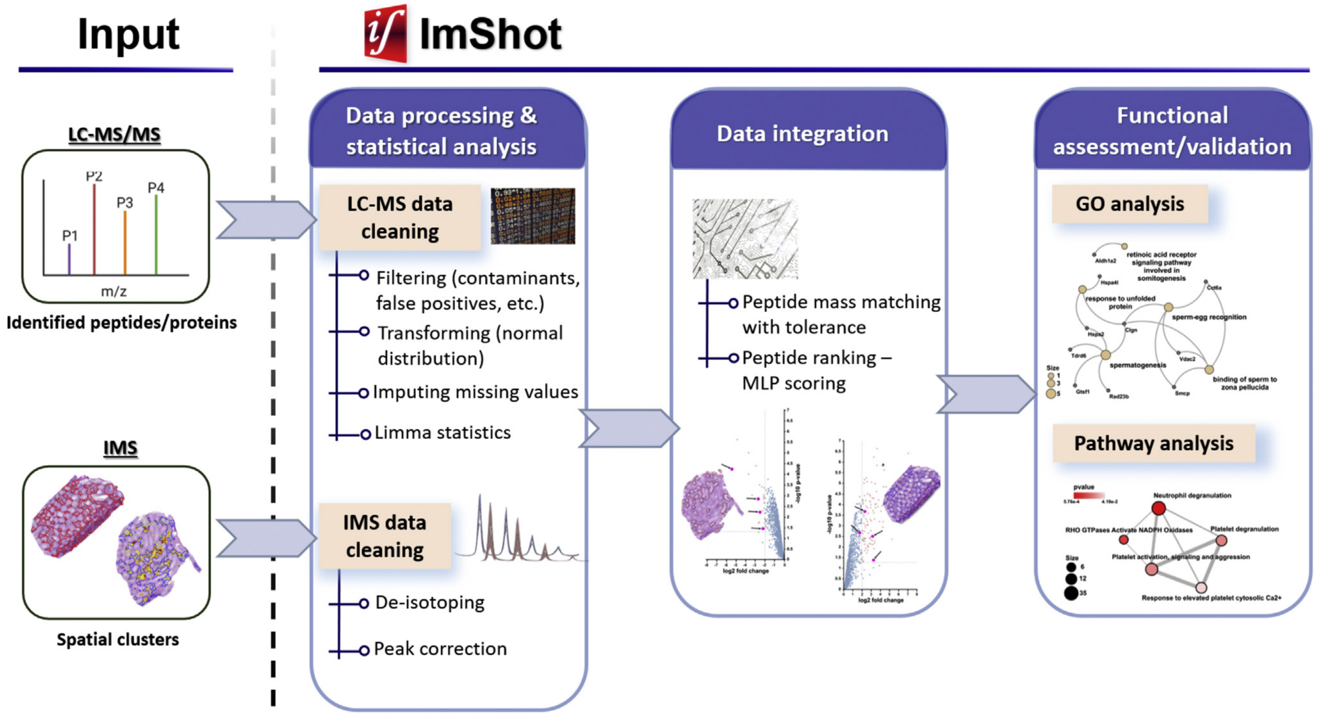

Input
ImShot
LC-MS/MS
P1
P2
P3
P4
m/z
Identified peptides/proteins
IMS
Spatial clusters
Data processing & statistical analysis
LC-MS data cleaning
Filtering (contaminants, false positives, etc.)
Transforming (normal distribution)
Imputing missing values
Limma statistics
IMS data cleaning
De-isotoping
Peak correction
Data integration
Peptide mass matching with tolerance
Peptide ranking – MLP scoring
log2 fold change
-log10 p-value
Functional assessment/validation
GO analysis
retinoic acid receptor signaling pathway involved in somitogenesis
response to unfolded protein
sperm-egg recognition
spermatogenesis
binding of sperm to zona pellucida
Size
Pathway analysis
pvalue
Neutrophil degranulation
RHO GTPases Activate NADPH Oxidases
Platelet degranulation
Platelet activation, signaling and aggression
Response to elevated platelet cytosolic Ca2+

Fig. 1. **ImShot modules and data integration pipeline**—Panel to the *left of dotted vertical line* (input) shows that ImShot accepts datasets from both the IMS (peptide clusters) and LC–MS (MaxQuant output: proteingroups.txt/peptides.txt) experiments as input. The Imshot panel consists of three segments. (i) Data processing and statistical analysis. This is responsible for transforming LC–MS and IMS datasets in a format that is compatible for data integration module. (ii) Data integration module. This segment identifies the parent protein of each IMS peptide by associating it to an LC–MS peptide based on mass matching and MLP scoring. (iii) Functional assessment/validation module. This serves as a validation tool for the MLP scoring by integrating information from the literature through GO and pathway enrichment analysis. The *arrows* show information flow between the modules to actualize the data integration pipeline. GO, Gene Ontology; IMS, imaging mass spectrometry; MLP, most likely peptide.

has options for removing exclusively enriched proteins, that is, proteins that are quantified exclusively in either of the groups. ImShot can further filter (optional) the dataset by removing proteins that were not quantified in at least *k* out of *N* replicates in each group. Multiple layers of filtering thus help significantly lower the impact of missing values on subsequent statistical analyses.

iii. The filtered dataset is log transformed to ensure it is normally distributed and so is suitable for subsequent statistical testing.

iv. The log transformation in the previous step produces many undefined values (NaNs) also popularly known as missing values. Missing values is a common scenario in proteomics that comprises of both, values missing at random and missing not at random. Missing at random can happen, for example, when either no peptide was found, or the peptide was misidentified. Missing not at randoms usually are more abundance dependent. Since such missing values interfere with the statistical tests, one needs to handle them appropriately. Although ImShot's aforementioned filtering modules greatly reduce the proportion of missing values, some proteins would still have them in at least one of the replicates leading to the NaNs. To deal with this scenario, the software uses a commonly used imputation algorithm assuming that proteins with low expression levels result in missing values. Such values are quite reliably estimated in ImShot by creating a tiny normal distribution by shrinking and downshifting the original data distribution toward low expression (supplemental Fig. S1) (35), from which the missing values are drawn at random. In addition to this imputation algorithm (*tiny*), ImShot also offers nine other frequently used imputation algorithms for maximum user flexibility (supplemental Table S1).

v. In the following step, data are normalized to deal with systematic shift in protein intensities that could arise from technical and other variations. The users can select between the following two modes of normalization depending on the type of data and biological system that they are using:

- Column-wise median normalization: In this mode, for each column (sample) in the data matrix (samples from both groups included), ImShot first computes the median difference (*i.e.*, the difference between measured intensities in a sample [column] and the mean of all rows/proteins) and then subtracts it from each row (protein) (36), the assumption being that most of the proteins do not change.
  In some specific situations, investigators may want to use the top *n* or some spike-in proteins to normalize the data. ImShot can also deal with that by allowing the users to select a set of proteins (from the rows of data matrix) as the spike-ins. However, this normalization feature that allows users to use *n* spike-in proteins (*n* < total number of proteins) is currently supported only by ImShot R package, and depending on user demands, it will be incorporated in the GUI of future version(s) of ImShot.
- Normalize by subtracting median: In this mode, ImShot normalizes the protein intensities in each sample by subtracting the median of the corresponding sample. Here, the samples are scaled so that they have the same median (zero).

Statistical Analysis Of Proteomics Datasets Using Moderated t *Test*—The processed data as mentioned in the preceding section are used for statistical analysis (two-group comparison test). To determine the significantly enriched proteins, ImShot uses Limma (linear models for microarray data) moderated *t* test statistics as suggested in the study by Kammers *et al.* (37). Limma was originally developed to find differentially enriched genes in microarray-based experiments, and for many years, it has been considered state of the art to analyze data from gene expression experiments such as RNA-Seq. It employs empirical Bayes approach that uses the entire dataset to shrink the estimated sample variances for each gene toward a pooled estimate (38, 39). This statistical approach results in much more stable and powerful inference compared with ordinary *t* statistics mainly when the number of replicates is small (39, 40). Very often, proteomics datasets come with small replicates/sample sizes where such Bayesian treatment is appropriate and has therefore gained some popularity within the proteomics community over time (41–45). ImShot employs the Limma R package (46) in the backend to computationally compare two groups (*i.e.*, healthy *versus* diseased) in proteomics datasets. For every protein, the *t* statistics t_{ord} is computed using the mathematical formula presented in (Equation 1). Where, lfc (log fold change) implies difference between means of the two groups in $\log_2$ scale and σ, $\sigma_{\mathrm{unscaled}}$ imply residual standard deviation and unscaled standard deviation, respectively.

$$t_{\mathrm{ord}} = \frac{lfc}{(\sigma_{\mathrm{unscaled}} {}^{*} \sigma)} \tag{1}$$

$$t_{\mathrm{mod}} = \frac{lfc}{\left(\sigma_{\mathrm{unscaled}} {}^{*} \sigma_{\mathrm{posterior}}\right)} \tag{2}$$

$$\Delta_{\sigma} = \frac{\sigma - \sigma_{\mathrm{posterior}}}{\sigma} {}^{*}100 \tag{3}$$

The Limma moderated *t* statistics t_{mod} is computed using the mathematical formula presented in (Equation 2). Where, $\sigma_{\mathrm{posterior}}$ implies posterior values of σ, which is learned by applying empirical Bayes method on full data. Therefore, if $\sigma > \sigma_{\mathrm{posterior}}$, then $|t_{\mathrm{ord}}| < |t_{\mathrm{mod}}|$.

Since a typical mass spectrometry experiment can identify and quantify hundreds to thousands of proteins, multiple comparison corrections are critical in controlling false positives. False positives are assessed in ImShot using *q* values (47, 48), which are defined as the lowest false discovery rate, at which a particular protein can be classified as differentially expressed (37). *q* Values are used in the same way as *p* values are used to control type I error. After multiple comparison corrections, if a protein is assigned a *q* value of 0.05, then one can anticipate that 5% of proteins with lower *p* values will be false positives. Although multiple comparison corrections are intended to reduce the false positives, adjusting for it might increase the number of false negatives or instances where an impact exists but is not detected as statistically significant (49). On such occasions when false negatives are prohibitively expensive, one may wish to avoid correcting for multiple comparisons. To account for this and hence allow higher user discretion, multiple comparison corrections are optional in ImShot.

Data Processing (MALDI–IMS)

The first step for processing IMS data involves the ever-challenging problem of deisotoping an IMS peptide spectrum. Owing to the lack of physicochemical separation of the peptides generated on tissues, a serious problem of overlapping isotopic envelopes arise in almost all the spectral files. Peaks at isotopic positions of one peptide are often masked by peaks belonging to entirely different peptide(s) in the tissue (Fig. 2*A*, *top panel*; spectrum in *.mis *files* section). Deisotoping of imaging mass spectra has therefore been an unresolved challenge in the field so far. A software module in ImShot (dedicated for deisotoping) generates monoisotopic IMS mass lists (Fig. 1) using the following procedure.

The algorithm begins with the assumption that distinct peptides (at isotopic positions) from a tissue display distinct distribution patterns (Fig. 2*A*, *top panel*; spectrum in *.mis *files* section). Therefore, in order

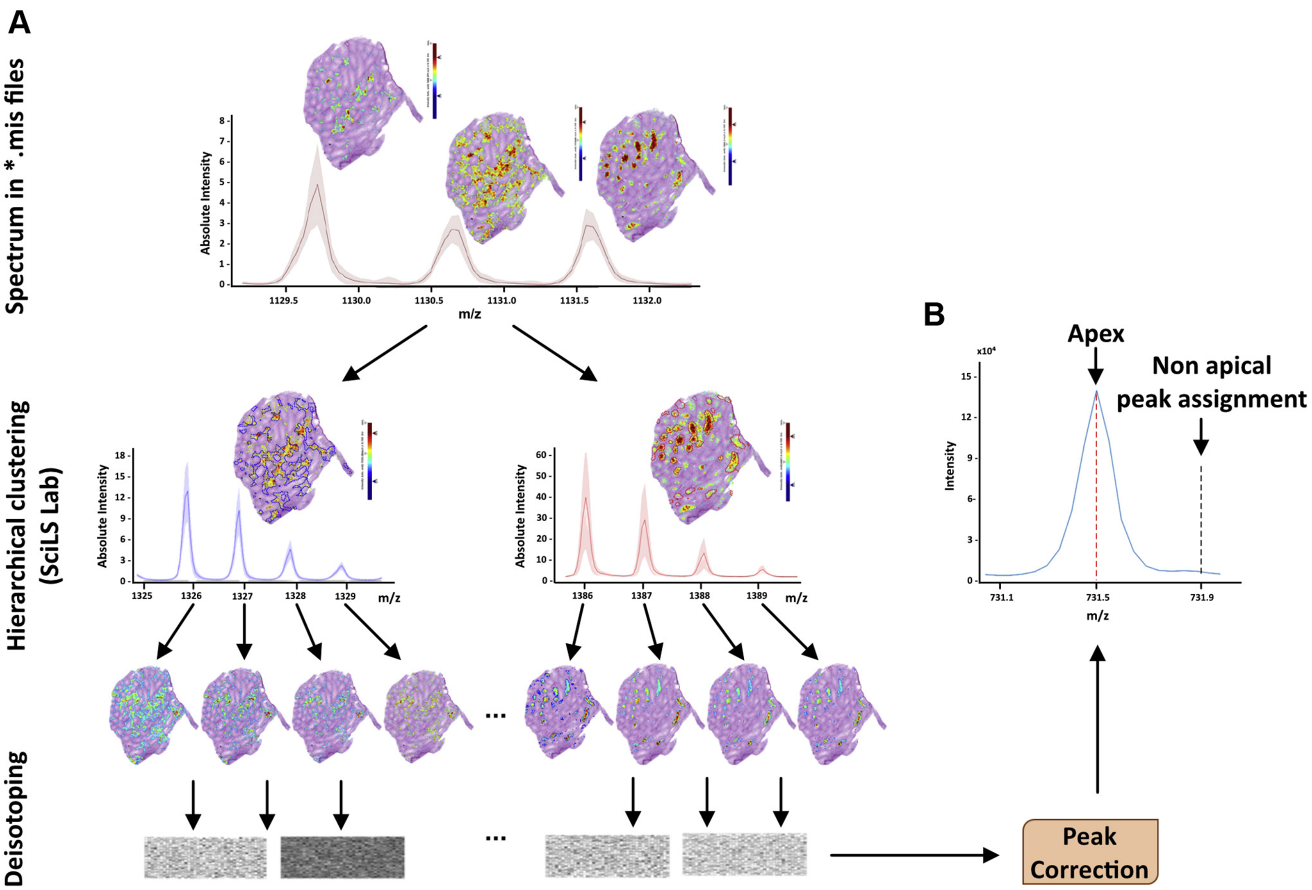


Fig. 2. **Generating monoisotopic IMS mass lists by filtering overlapping spectra.** *A*, *upper panel*, (spectrum in *.mis files) shows a supposed isotopic envelope extracted from a MALDI–IMS spectrum. The supposed isotopic peaks show very distinct spatial distribution patterns (HE tissue images alongside the peaks). The *middle panel* (hierarchical clustering [SCiLS Lab]) shows the segregation of the complex spectra into those of distinct peptides (m/z = 1326 and 1385.75 in this case). The spatial distributions of masses at isotopic positions are identical in this case (ion images below the isotopic envelopes). Deisotoping algorithm is applied on these spectra to generate the monoisotopic IMS mass lists (*bottom panel*). *B*, nonapical peak assignments in some cases by proprietary software, for example, SCiLS lab. It highlights the importance of the peak correction module in ImShot. IMS, imaging mass spectrometry.

to parse the tissue into maps of distinct peptides, we have applied the unbiased hierarchical clustering algorithm of SCiLS Lab (SCiLS, www.scils.de) (33) on the entire IMS dataset (Fig. 2*A*, *middle panel*; *Hierarchical clustering* section). Since spatial distributions of isotopic peaks of the same peptide are supposed to be identical, we applied the deisotoping algorithm on the mass lists that distinguished one cluster from the other (Fig. 2*A*, *lower panel*; *Deisotoping* section). We did not encounter any isotopic envelope violating the aforementioned condition in the example IMS dataset (7).

Deisotoping was performed using standard tolerances, which for MALDI-TOF peptides were ±0.15 and 50% for m/z and intensity, respectively (50), for example, for any $m/z = m$, we removed all the m/z (m_i) values falling within the interval: $m + 0.85 \geq m_i \geq m + 1.15$. These values can be adjusted in case of high-resolution IMS measurements to account for monoisotopic peaks that might be present in the same spatial cluster and within the isotopic mass tolerances.

Following the deisotoping step, we occasionally observed nonapical value assignment of a small fraction (~20–25%) of all the monoisotopic m/z peaks in a cluster (Fig. 2*B*). For example, in case of an apex of a peptide peak at m/z = 731.5, SCiLS Lab (33) assigns the value at m/z = 731.9. This happens mainly because it deals with m/z intervals rather than m/z peaks. Since our aim is to "identify" peptides from IMS by comparing it with LC–MS data, we applied a peak correction algorithm to the "incorrectly" assigned m/z values (Fig. 2*B*). To correct a peak corresponding to an m/z value (say m), ImShot scans a window of m/z +1 that contains m. If the intensity of m is not the highest within that window, then it updates m with the m/z value corresponding to the highest value there. As the ions are mostly singly charged in MALDI, this strategy ensures that the correct apex value is selected. However, in case of high-resolution IMS datasets, there is a rare possibility that peaks with a nonapical value occurred within the aforementioned window. Based on user input and occurrence analysis, this particular situation would be dealt with in the next version of ImShot.

The processed mass list from the aforementioned step is then used for data integration with LC–MS data and identification of parent proteins using user-defined mass matching and MLP scoring.

Development of ImShot R Package

The ImShot R package has been developed to offer greater flexibility to the users. It has five main functions to implement IMS and LC–MS data integration pipeline and also to handle LC–MS data separately. The main function(s) are further supported by many subfunctions rendering the source code of ImShot entirely modular. Following are the five main functions of the ImShot R package:

i. two_grp_limma—This function filters an LC–MS dataset (proteingroups.txt/peptides.txt from MQ output), imputes missing values, normalizes, performs two group comparisons using Limma statistics, and writes the results in the user config directory according to the aforementioned steps.

ii. Deisotope_masslist—This function desisotopes IMS masslists and corrects *m/z* peaks for mass deviation. Results are written in a timestamped folder inside user config directory depending on the users' OS as described in the section *Testing ImShot desktop app* in the supplemental data.

iii. ims_lcms_integration—As the name suggests, this function implements data integration by combining outputs from two_grp_limma and deisotope_masslist functions with the help of MLP scoring followed by writing the results in the user config directory as described in the section *Testing ImShot desktop app* in the supplemental data.

iv. go_enrichment_analysis—This function performs GO enrichment analysis using ClusterProfiler R package (51). However, one node label frequently overlaps with another in the network plot created by the ClusterProfiler package, making the plot difficult to interpret. Furthermore, the ClusterProfiler package does not allow users to modify the (*x*,*y*) coordinates of nodes to improve clarity. To overcome the challenge, this ImShot function facilitates plotting results directly on a Cytoscape (https://cytoscape.org/development_team.html) (52) session, where users can choose from a wide range of available toolboxes to customize their network graphs. It is important to note that Cytoscape must be running before executing this script. Finally, this function writes the results after enrichment analysis in the user config directory as described previously.

v. pathway_enrichment_analysis—This function performs pathway enrichment analysis using ReactomePA R package (53). Like the go_enrichment_analysis function, this function too facilitates plotting results directly on a Cytoscape session to bypass the issues described earlier. Cytoscape must be running before executing this script, and it also writes the results after enrichment analysis in the user config directory in the aforementioned manner.

The help functionality of R can be used very effectively by the user to know about the operational details of the functions. For example, to know more about *two_grp_limma()*, the users can type ?ImShot::*two_grp_limma* in RStudio console. Advanced users can also freely modify the functions to implement desired functionalities or to expand the package's functionalities by augmenting it with new functions.

Development of ImShot Desktop Application and GUI

The ImShot desktop application has been developed using several programming languages. The front end is written using hypertext markup language (HTML), cascading style sheets (CSSs), JavaScript (JS), and the backend is mostly R (Fig. 3, *left panel*). JS also participates in the backend computation by creating appropriate data structures for plotting the data. We used the child_process module of node.js to call functions written in R programming language that perform data wrangling in the backend (Fig. 3, *left and middle panels*). We used the child_process.spawnSync function, which spawns child process in a synchronous manner that blocks the event loop until the spawned process either exits or is terminated. The data are then passed from R environment to JS in the form of a JS Object Notation (JSON) object https://www.json.org/json-en.html; (Accessed on 2021/07/02) (Fig. 3, *left panel*) that optimizes communication and editing at every level. ImShot JSONifies (encodes in JSON format) the data to be sent to R, and in R environment, it gets un-JSONified (decoded from JSON format) so that R code can use them. Finally, the software JSONifies the R output and sends it to JS, where it gets un-JSONified for displaying in the front end (Fig. 3, *left panel*). Thus, JS facilitates the communication between front ends and back ends by acting as an interface between them.

ImShot then employs the open-source software framework Electron that enables building desktop applications by integrating web technologies, such as JS, HTML, and CSS. It does so by combining Chromium rendering engine and the Node.js runtime. Figure 3, *middle panel*, depicts the multiprocess architecture of Electron. The main process's task is to start the application and respond to its lifecycle events such as creation and destruction of renderer process. It is also responsible for communicating with OS *via* system application programming interfaces (APIs) (Fig. 3, *right panel*). The Renderer process uses Chromium engine to render a web page as an independent process. It handles fetching and rendering HTML, loading any referenced CSS and JS, styling the page accordingly, and executing the JS. The Node.js runtime uses Google's open source V8 engine to interpret JS and provide APIs for accessing the file system, loading code from external modules, and communicating with other programming languages.

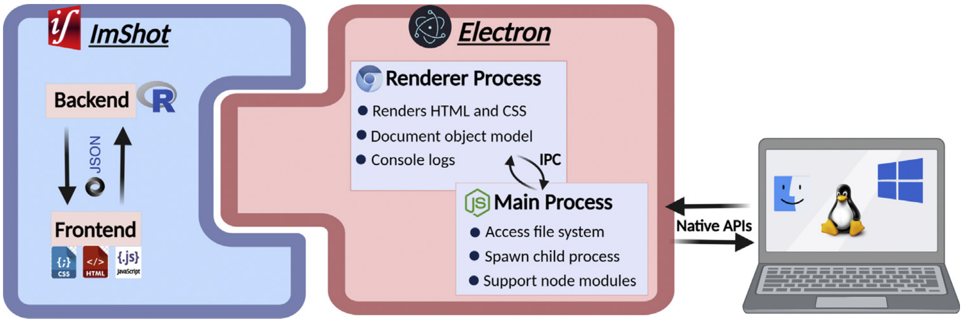


FIG. 3. **Software architecture.** *Left panel (blue)*, shows the architecture of ImShot desktop application, which comprises of two parts: front end and back end. Front end is developed using HTML, CSS, and JavaScript, and the backend is programmed in R. The two ends communicate with each other by passing data in the form of JSON objects. Front ends and back ends are packaged using Electron framework that has a multiprocess architecture (*middle panel*, *red*). The renderer process is mainly responsible for rendering the GUI, and the main process manages the state of the application by communicating with operating system *via* native application programming interfaces (APIs) (*right panel*). The Renderer and main processes communicate with each other through interprocess communication (IPC), which is critical for successfully executing user requests. Created with BioRender.com. CSS, cascading style sheet; GUI, graphic user interface; HTML, hypertext markup language; JSON, JavaScript Object Notation.

RESULTS

Moderated t *Test Yields More Significant and Biologically Relevant Proteins*

Application of Limma-based moderated *t* test on the dataset from our recent study on aromatase-induced male infertility (7) shows that this form of analysis yields more statistically significant proteins (Fig. 4, *green points* on the volcano plot; Table 1) as compared with the ordinary *t* test. For all these proteins, we observe that the percentage shrinkage (Δ_σ) in sample variance (computed using (Equation 3)) is always positive (Table 1) and $|t_{\mathrm{mod}}|$ is always greater than $|t_{\mathrm{ord}}|$. As higher *t* value is associated with smaller *p* value, we observe that Bayesian modeling yields more statistically significant proteins when compared with ordinary *t* test by shrinking the sample variance toward a pooled estimate.

However, to learn if these additional statistically significant proteins are at all relevant biologically, we performed GO enrichment analysis using the GO analysis module of ImShot. In GO-cellular component enrichment network, we notice that most of the proteins that are significantly upregulated upon aromatase overexpression are involved in regulation of the extracellular matrix (ECM) (Fig. 4, *lower red panel*). Interestingly, it has been shown that components of the ECM are upregulated in men suffering from infertility (54, 55). In case of WT, we observe the prominence of acrosomal membrane and protein complexes required for high-energy cellular processes (as expected in case of normal spermatogenesis) (Fig. 4, *lower green panel*). Together, these results justify the application of Bayesian modeling over ordinary *t* test: we not only get more proteins that are statistically significantly different between two conditions but also can get more information that describes the condition biologically.

"Identification" of Proteins In Situ *Through Data Integration and Peptide Ranking*

Results from the deisotoping module are used as the IMS input for data integration. Our example dataset revealed a substantial reduction in the number of peaks after deisotoping (Table 2). Since the isotopic peaks of the same peptide have identical spatial distribution, deisotoping could get rid of false-positive peaks. Monoisotopic mass lists from the IMS experiments are compared with LC–MS data to identify the corresponding parent proteins in the data integration modules (Fig. 5, *left and right panels*). First, the monoisotopic mass list for every spatial cluster is searched within either diseased or healthy set of enriched LC–MS peptides depending on the prevalence of the respective cluster (Fig. 5, *left panel*). Masses that distinguish a diseased peptide cluster from a healthy one or vice versa are thereby selected for the intended differential analysis. Since the accuracy of measurement differs according to the measurement platform (ion source, mass analyzer, etc), the search is performed within a certain tolerance (τ). This part of the module has been kept flexible (user-specified input) (supplemental Fig. S2, snap from the GUI) keeping in mind the wide variety of measurement platforms that the users might use.

Owing to the relatively low accuracy of IMS as compared with conventional shotgun proteomic measurements, the tolerance search potentially yields one-to-many mapping between IMS and LC–MS peptides, that is, one IMS peptide mass is annotated to multiple LC–MS peptides originating from different parent proteins (Fig. 5, *right panel*). Results of the tolerance search on our example data show that ~63% of *m/z* values (spanning over the IMS clusters mentioned in Table 2) bear the one:many correspondences with the identified peptides in LC–MS. To resolve this ambiguity, we devised a novel scoring method that ranks the identity of peptides (as being part of the parent protein) based on the following equation:

$$\mathrm{MLP} = \frac{\mu * \log_2 fc}{p_{\mathrm{mod}}} \tag{4}$$

where μ is the mean intensity (after log transformation) of a peptide across the replicates in either diseased or healthy group, p_{mod} is the Limma moderated *p* value of the same peptide, and $\log_2 fc$ implies the fold change between the diseased and healthy groups, which is defined as follows,

$$\log_2 fc = \mu_{\mathrm{diseased}} - \mu_{\mathrm{healthy}} \tag{5}$$

Likelihood of a peptide to belong to its corresponding identified parent protein was correlated to increasing MLP score for that peptide based on the following reasonings:

i. Peptides of relatively higher abundance are preferably detected in MALDI–IMS, mainly because of the lack of any separation technique and competitive cocrystallization of matrix/biomolecules (μ). Hence, peptides (belonging to certain proteins) having a higher μ value among the multiple possibilities are most likely the ones that are detected in IMS.

ii. The search space for a peptide belonging to a cluster detected in IMS measurements is narrowed down to either healthy or diseased LC–MS data depending on their occurrence in the corresponding tissues ($\log_2 fc$). This increases the likelihood of a peptide belonging to a particular protein with very high confidence.

iii. Inclusion of the moderated *p* value in the scoring system is used to increase the likelihood of a peptide belonging to a given parent protein even further (p_{mod}). Using *p* values arising from Limma statistical tests increases the search space further. Lower the p_{mod}, higher the score and higher the probability of an IMS peptide to belong to its corresponding identified protein.

Therefore, peptides from spatial IMS clusters with top MLP scores are regarded as belonging to the corresponding parent protein identified in LC–MS (Fig. 5, *right panel*). This workflow is designed to focus on the most differentially expressed

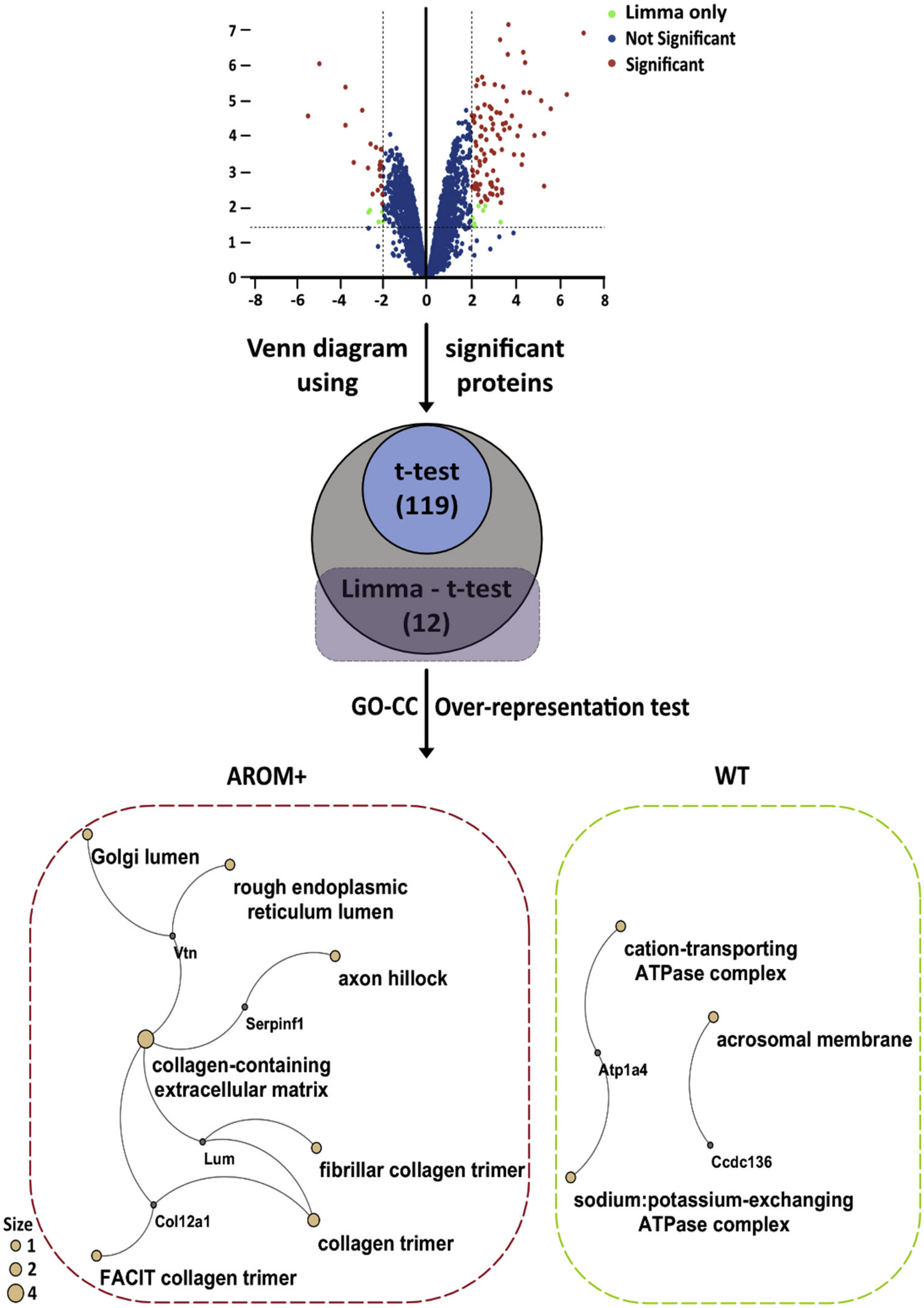


FIG. 4. **Limma-moderated *t* test provides more powerful inference.** *Top panel* shows a volcano plot after two-group comparison test using Limma statistics where *dots* with *red colors* correspond to statistically significant proteins having ($lfc > 2 \parallel lfc < -2$) and $p < 0.05$ that are overlapping between standard *t* test and Limma. The *green dots* in the plot show the proteins that become statistically significant only when Limma is applied. The Venn diagram in the *middle panel* demonstrates that Limma statistics yields 12 more proteins than ordinary *t* test. These were used in GO-CC enrichment analysis whose results are depicted in the *bottom panel* in the form of a gene–GO term network. Two distinct GO clusters are observed: the *red rounded dashed rectangle* displays the cluster of terms enriched in AROM+ proteins, whereas the *green rounded dashed rectangle* highlights the cluster of terms enriched in WT proteins. *Tiny fixed sized gray nodes* in the network represent genes, and *larger light-colored nodes* (variable sizes, see size legend) represent GO terms. An edge between a gene node and GO term node indicate that the term was not enriched by chance. CC, cellular component; GO, Gene Ontology.

TABLE 1
Statistically significant proteins determined exclusively using Limma statistics

Gene	lfc	t_{ord}	p_{ord}	t_{mod}	p_{mod}	σ	$\sigma_{\mathrm{posterior}}$	Δ_{σ}
Fsip2	−2.58786	−2.54893	0.063377	−3.17401	0.017473	1.243452	0.998568732	19.69
Afm	2.304437	2.729307	0.05248	3.371293	0.013497	1.034089	0.837170525	19.04
Vtn	2.282418	2.318201	0.081297	2.88337	0.025814	1.20584	0.969483584	19.6
Hist1h1e	2.10801	2.072954	0.106872	2.581464	0.039149	1.245457	1.000120384	19.7
Lum	2.611408	2.718441	0.053072	3.377887	0.013382	1.176523	0.946837126	19.52
Serpinf1	2.453236	2.521249	0.06527	3.134472	0.018413	1.191706	0.958562629	19.56
Ccdc136	−2.11621	−2.22431	0.090179	−2.7628	0.030447	1.165223	0.938113925	19.49
Col12a1	3.304637	2.097001	0.104005	2.639595	0.036104	1.93006	1.533317489	20.56
Lypd4	−2.61134	−2.53249	0.064494	−3.1553	0.017911	1.262879	1.01360433	19.74
Fmo2	2.138849	1.976454	0.119285	2.466548	0.045988	1.325376	1.062028449	19.87
Tex33	−2.00396	−2.1484	0.098158	−2.66628	0.034791	1.142407	0.920511851	19.42
Atp1a4	−2.0369	−2.51776	0.065513	−3.1028	0.019206	0.990835	0.804009894	18.86

lfc, log fold change; σ, sample standard deviations for each gene/protein; $\sigma_{\mathrm{posterior}}$, posterior values for σ; $\Delta\sigma$, percentage shrinkage; p_{ord}, p values corresponding to the t-statistics (t_{ord}); p_{mod}, p values corresponding to the moderated t statistics (t_{mod}).

peptides under an altered physiological state. The input IMS data ensure selection of peptides distinguishing the healthy from the diseased state. Further emphasis on the higher fold change and lower p values in the peptide ranking step reinforces the focus on proteins discriminating a diseased state from a healthy one.

Validation of "Identified" Proteins in ImShot

Additional validation of the proteins with top MLP scores is required to impart further confidence in the applicability of this scoring method, in general. In addition to validating experimentally the distribution pattern of a subset of peptides identified in IMS measurements (7), we attempted to further validate the scoring here by using a combinatorial approach involving modules from ImShot itself, available public data, and correlation analysis of multiple peptides assigned to the same proteins within a cluster.

TABLE 2
Number of peaks (peptide masses) in IMS clusters before and after data cleaning (deisotoping + peak correction)

Cluster	Before	After
WT_Tubular_cluster_1	110	51
WT_Tubular_cluster_2	4	3
WT_Tubular_cluster_3	235	108
WT_Tubular_cluster_4	244	114
WT_Tubular_cluster_5	116	55
WT_Tubular_cluster_6	37	27
WT_Tubular_cluster_7	183	135
WT_Tubular_cluster_8	2	2
WT_Tubular_cluster_9	11	6
WT_Tubular_cluster_10	9	7
WT_Tubular_cluster_11	113	54
WT_Tubular_cluster_12	14	7
WT_Tubular_cluster_13	154	78
WT_Tubular_cluster_14	2	1
AROM_Interstitial_cluster	40	37
Total masses	1274	685

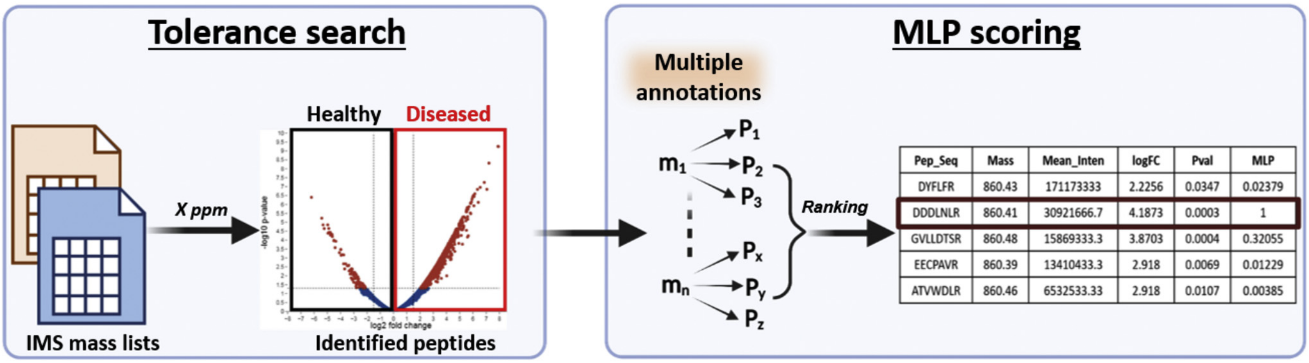
Computational Validation Through ImShot Modules

The GO and pathway analysis modules of the functional assessment/validation section (Fig. 1, *right panel*) are used here to assess the functional relevance of the identified peptides according to their spatial localization. We first opted to validate the parent proteins of peptides (with highest MLP score) from a cluster observed exclusively in the interstitial spaces of AROM+ testis (Fig. 6*A*).

GO Analysis—This module allows users to associate a common theme to the genes/proteins of interest that can help answer the biological question. GO provides annotation for genes or gene products at different domains: cellular component, molecular function, and BPs that are organized in the form of directed acyclic graph data structure. It is possible that proteins could be annotated to multiple GO nodes. Moreover, because of the nature of directed acyclic graph data structure, a gene annotated to a particular node also inherits annotation from the ancestors of that node. Therefore, in order to find out if a GO term is enriched in specified list of genes not by chance, ImShot calculates p values as proposed in the study by Boyle *et al*. (56):

$$p = 1 - \sum_{i=0}^{k-1} \frac{\binom{M}{i} * \binom{N-M}{n-i}}{\binom{N}{n}} \tag{6}$$

Where, N is the number of genes/proteins in background list, M is the number of genes within that list that have direct/indirect annotation to the GO node of interest, n represents the length of the list corresponding to the genes of interest, and k is the number of genes within that list, which are annotated to the node. ImShot uses an R package called ClusterProfiler in the backend to perform the GO over-representation test. For the background gene set, ImShot allows the user to either use the global background provided in ClusterProfiler or a gene/protein list of their own, which contains customized background list for user-specific needs. The results are displayed in the form of a network graph



Pep_Seq	Mass	Mean_Inten	logFC	Pval	MLP
DYFLFR	860.43	171173333	2.2256	0.0347	0.02379
DDDLNLR	860.41	30921666.7	4.1873	0.0003	1
GVLLDTSR	860.48	15869333.3	3.8703	0.0004	0.32055
EECPAVR	860.39	13410433.3	2.918	0.0069	0.01229
ATVWDLR	860.46	6532533.33	2.918	0.0107	0.00385

FIG. 5. **Data integration challenges and solutions.** *Left panel*, (tolerance search) illustrates the concept used in the tolerance search module of ImShot. IMS mass lists from one cluster of any group (healthy/diseased) is searched inside LC–MS mass list (volcano plot) of the corresponding group within a user-defined tolerance (*X ppm*). Because of the difference in accuracy of measurement between the two orthogonal MS platforms, search results are ambiguous, that is, multiple LC–MS peptides can be annotated to an IMS peptide as illustrated in the *right panel*. The data integration module of ImShot resolves this ambiguity by ranking LC–MS peptides by MLP scores where the peptide with highest MLP score is considered to be the most likely one (*right panel*). Created with BioRender.com. IMS, imaging mass spectrometry; MLP, most likely peptide.

that shows association between gene and GO terms, and an edge is drawn between a gene and a GO term if gene is found enriched in that term (Fig. 6*C*).

We observe that peptides from proteins like mimecan, different chains of collagen, and prolargin, which have previously been shown to be involved in ECM assembly and regulation (7, 55, 57), have acquired the highest MLP scores (Fig. 6*A*). In addition, we find that the interstitial cluster is enriched in BPs related to connective tissue formation involving ECM components (Fig. 6*C*) and hemostasis. This is in line with the observation that interstitial spaces and the ECM components involved therein are severely affected in the AROM+ phenotype (54, 55). At the same time, these proteins are also observed to be highly enriched in AROM+ LC–MS data (Fig. 6*B*). Applying expert knowledge and database mining on the peptides identified with highest, second highest, and third highest MLP scores for the aforementioned cluster (Fig. 6*A*), we observed a decrease in biological relevance with lower MLP scores. Therefore, the MLP scoring–based ranking method is providing us with probable protein identification *in situ* with reasonable accuracy and minimum false positives.

Pathway Analysis

Pathway analysis module allows users to associate a common theme to the genes/proteins of interest by annotating them to the biological pathways. Often knowledge of affected biological pathways can help answer the biological question. Here, we used the R package ReactomePA (53) to discover biological pathways in which the genes/proteins of interest participate. Like the GO analysis module, the ReactomePA package uses hypergeometric distribution model to calculate *p* values to determine whether any pathways in Reactome database occur in a specified list of genes at a frequency greater than that would be expected by chance. In addition, ImShot also supports pathway enrichment analysis using the Kyoto Encyclopedia of Genes and Genomes database, thereby increasing the applicability of the software for the community. The results are displayed in a similar way as the GO results (Fig. 6*D*).

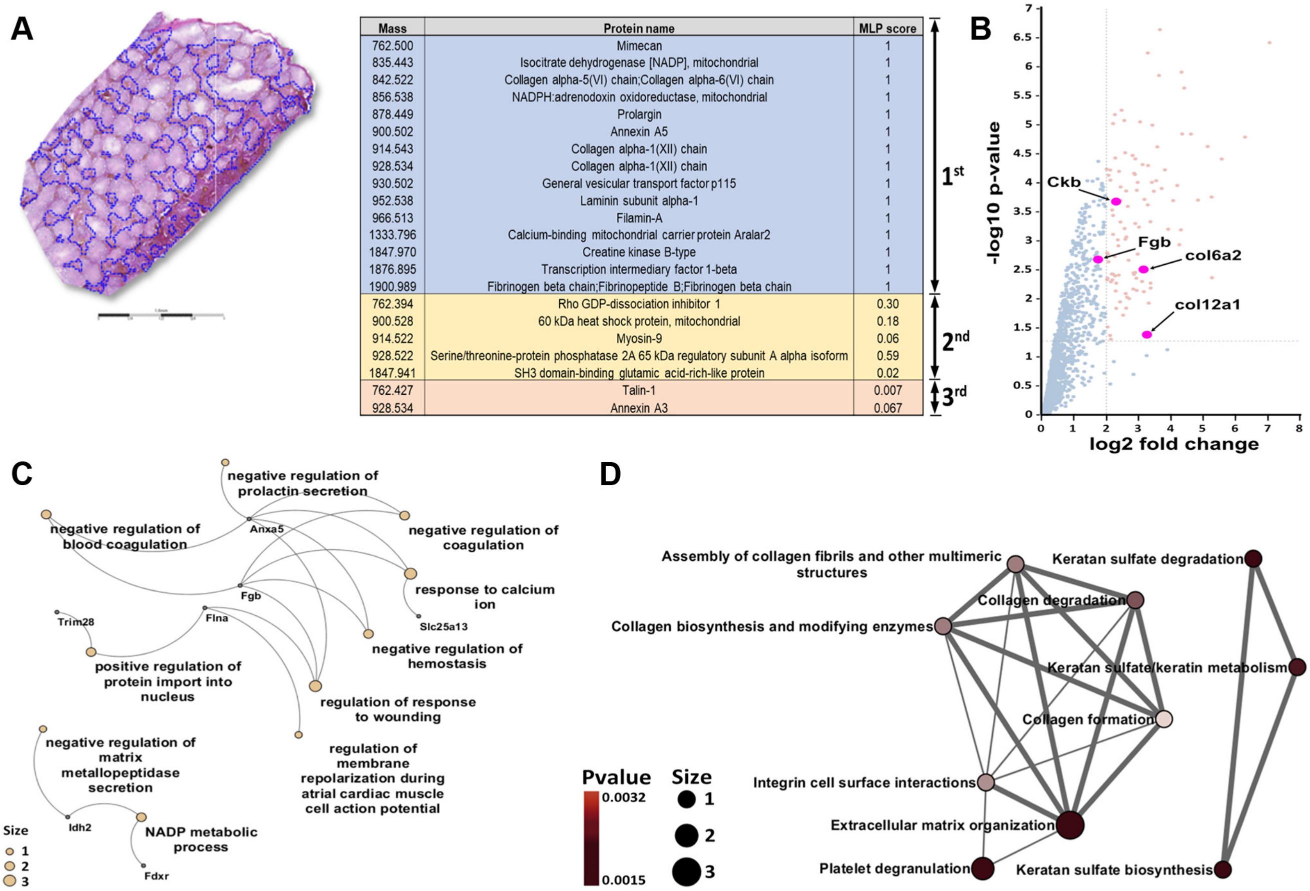
Consistent with the GO analysis and other findings, we see that ECM organization through collagen synthesis and assembly (Fig. 6*D*) and immune responses characterize the interstitial cluster in AROM+ testis. This further supports the validity of our ranking method for identifying peptides *in situ*.

Validation Based on Public Data

We have primarily used publicly available data for validating the *in situ* location of the proteins identified by ImShot. Two peptide clusters from the two different biological conditions (WT and AROM+) were selected that represented two distinct testicular tissue structures (seminiferous tubules and interstitium) (Fig. 7*A*) for more robust validation. The list of identified proteins (UniProt Ids) from each cluster (top MLP scores) was searched for their expression data in the mouse gene expression database (http://www.informatics.jax.org/expression.shtml) using the following workflow:

i. First, we navigated to the “search with gene list” page (http://www.informatics.jax.org/gxd/batchSearch) and pasted the UniProt Ids in the ID/Symbols field and filtered the “Search by” field to “UniProt ID” followed by the search function.
ii. Following the search results, we analyzed the “Tissue x Gene Matrix” tab after filtering the results for Anatomical System = Reproductive system and Theiler Stage = TS28 (postnatal).
iii. The resulting matrix was further filtered sequentially for male reproductive system, testis, interstitium of the testis, and seminiferous tubule.

The proteins that could not be matched to any mouse expression data as specified previously were searched for



Mass	Protein name	MLP score
762.500	Mimecan	1
835.443	Isocitrate dehydrogenase [NADP], mitochondrial	1
842.522	Collagen alpha-5(VI) chain;Collagen alpha-6(VI) chain	1
856.538	NADPH:adrenodoxin oxidoreductase, mitochondrial	1
878.449	Prolargin	1
900.502	Annexin A5	1
914.543	Collagen alpha-1(XII) chain	1
928.534	Collagen alpha-1(XII) chain	1
930.502	General vesicular transport factor p115	1
952.538	Laminin subunit alpha-1	1
966.513	Filamin-A	1
1333.796	Calcium-binding mitochondrial carrier protein Aralar2	1
1847.970	Creatine kinase B-type	1
1876.895	Transcription intermediary factor 1-beta	1
1900.989	Fibrinogen beta chain;Fibrinopeptide B;Fibrinogen beta chain	1
762.394	Rho GDP-dissociation inhibitor 1	0.30
900.528	60 kDa heat shock protein, mitochondrial	0.18
914.522	Myosin-9	0.06
928.522	Serine/threonine-protein phosphatase 2A 65 kDa regulatory subunit A alpha isoform	0.59
1847.941	SH3 domain-binding glutamic acid-rich-like protein	0.02
762.427	Talin-1	0.007
928.534	Annexin A3	0.067

FIG. 6. **Validating MLP scoring computationally.** *A*, HE-stained image of AROM+ mouse testis. The *deep blue pattern* within the tissue represents an interstitial cluster detected exclusively in AROM+. The peptides from this cluster were searched in corresponding LC–MS data, and the results after MLP scoring are shown in different colors according to the ranks. *B*, some proteins having peptides with top MLP scores are highlighted (in *pink*) in the volcano plot, showing that they were also highly enriched in the LC–MS data corresponding to AROM+ mice. *C*, gene–GO term network after over-representation test using proteins from the table annotated with highest MLP scores (first group). *Tiny fixed sized gray nodes* in the network represent genes, and *larger light-colored nodes* (variable sizes, see size legend) represent GO terms. An edge between a gene node and GO term node indicates that the term was not enriched by chance. *D*, gene–pathway network after over-representation test using proteins from the table annotated with highest MLP scores (first group). Nodes in the network represent Reactome pathways. Two pathways are joined with an edge if they share enriched (not by chance) genes. Thickness of an edge is proportional to the number of common genes. Nodes are colored according to *p* value of over-representation test, and the color gets darker as the *p* value decreases. GO, Gene Ontology; MLP, most likely peptide.

their corresponding human orthologs in the Human Protein Atlas (58) https://www.proteinatlas.org/humanproteome/tissue; (Accessed on 2021/04/24) by searching with the corresponding gene names in the search tab of the website. The results were further filtered by looking specifically into the "Tissue" section followed by selecting for "Male tissues" and "Testis." In parallel to the aformentioned data mining workflows, localization data in testis for the identified proteins were also looked for in general scientific literature using the PubMed search engine (https://pubmed.ncbi.nlm.nih.gov/) (supplemental files S1 and S2).

All the 41 proteins identified in the representative peptide cluster from seminiferous tubules (supplemental file S1) showed tubular localization when analyzed for evidence in mouse testis (Fig. 7*A*). In case of the interstitial cluster that number was 7 of 15 proteins (Fig. 7*A*). Since the AROM+ is a specific disease model, it is plausible that there are certain proteins that are upregulated in the interstitium, which was otherwise not observed under physiological conditions. We therefore looked for association of the MLP-identified interstitial proteins with ECM, components and factors of which are highly upregulated in the interstitium of the AROM+ (supplemental file S2). Using this exhaustive validation approach, we see that ImShot could successfully identify most of the proteins (false positive rate of 3.5%) in the two representative testis clusters.

Correlation Analysis

In the representative peptide clusters, we further looked for multiple peptides that were assigned to the same protein and

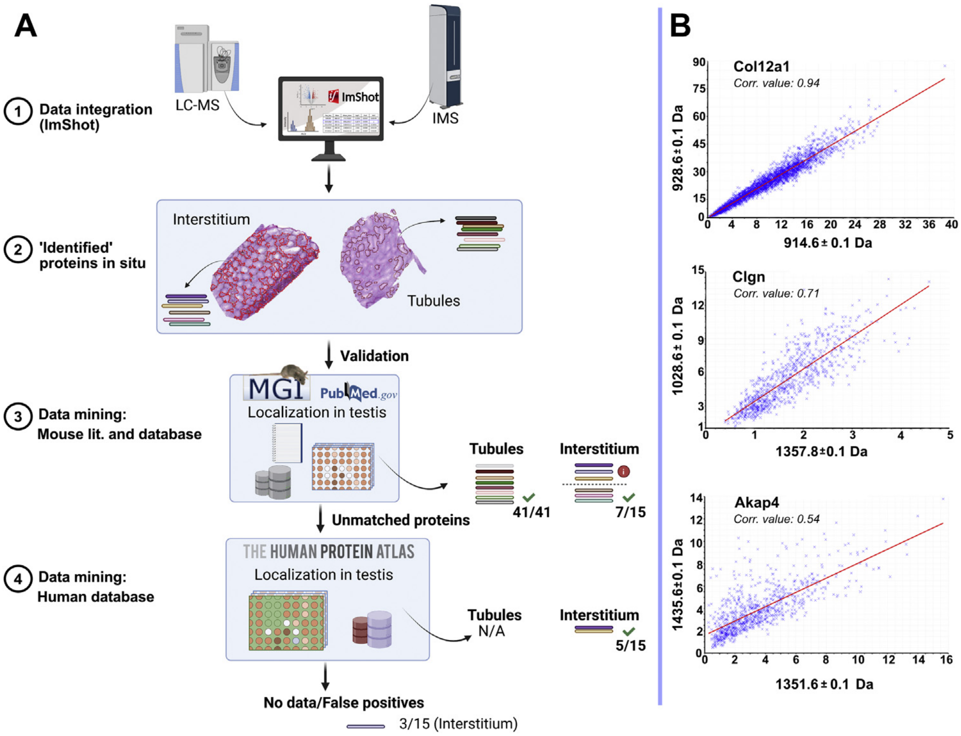


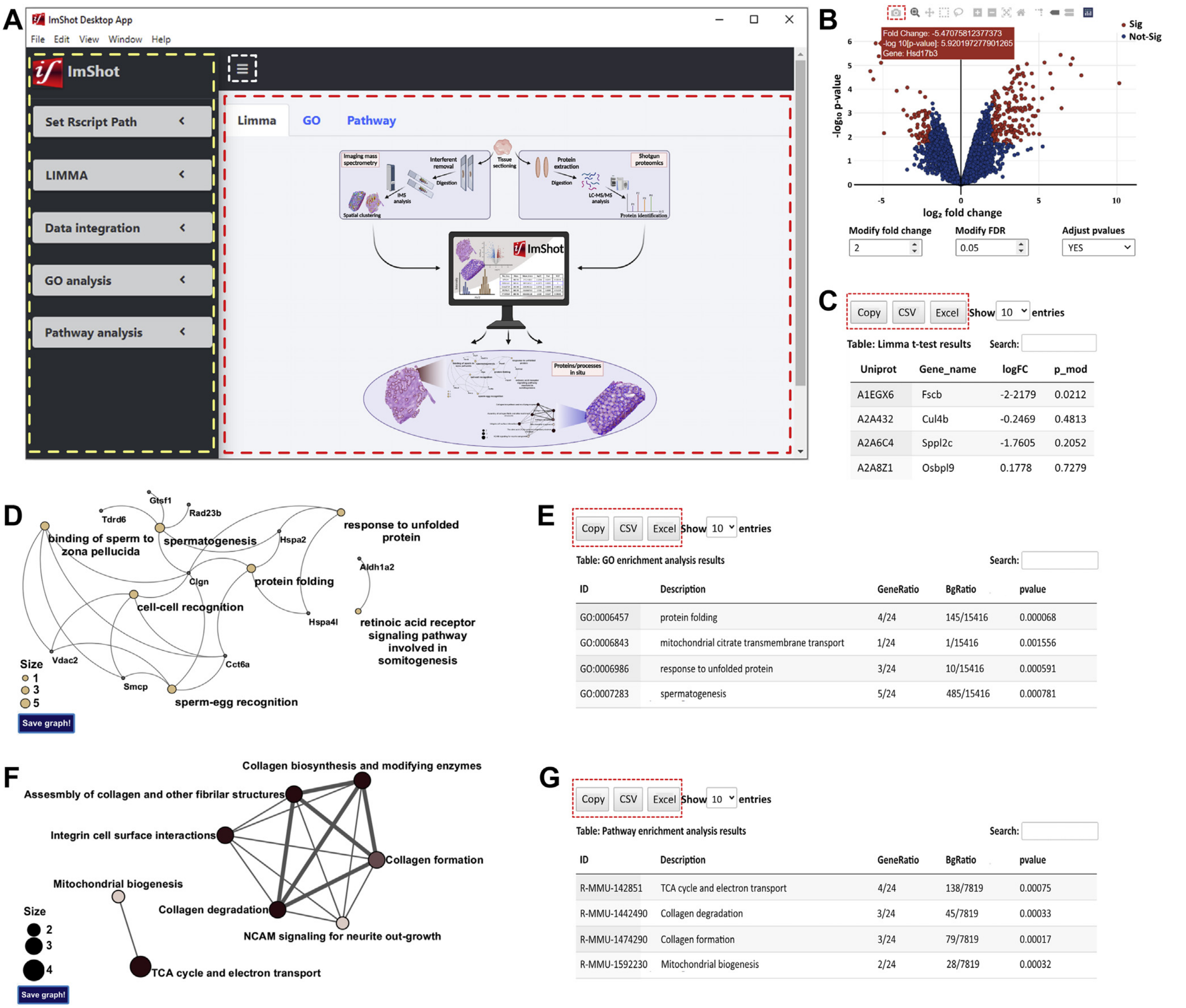
FIG. 7. **Validation of ImShot identified proteins.** *A*, corroborating the tissue localization of MLP identified proteins (panels 1 & 2; supplemental files S1 and S2) in the seminiferous tubules or interstitial space. Data mining approaches revealed tubular localization (in mouse testis, (68)) of all the 41 proteins identified in the tubular cluster (panel 3; supplemental file S1). In case of the interstitial cluster of AROM+, seven proteins were observed to be located in mouse testis interstitium and/or associated with the extracellular matrix (panel 3). Further mining in the Human Protein Atlas (panel 4, (58) https://www.proteinatlas.org/humanproteome/tissue; (Accessed on 2021/04/24)) confirmed the interstitial localization of five more proteins (human orthologs) in the interstitium of the testis. Of the remaining three proteins, two were possible false positives, and any localization data were not available for one (supplemental file S2). *B*, spatial correlation analysis of multiple peptides assigned to the same protein by ImShot (for tissue distribution pattern, see supplemental Fig. S11). *Top*, two peptides annotated to collagen alpha-1 (XII) chain (Col12a1) in the interstitium of AROM+. *Middle*, two peptides annotated to Calmegin (Clgn) in the seminiferous tubule of WT. *Bottom*, two peptides annotated to A-kinase anchor protein 4 (Akap4)in the seminiferous tubule of WT. All the correlation analyses were carried out within the respective IMS clusters by the correlation function of SCiLS Lab (33). Refer to supplemental Fig. S11 for the distribution patterns of Col12a1, Clgn, and Akap4. IMS, imaging mass spectrometry; MLP, most likely peptide.

performed a spatial correlation analysis on the IMS dataset (Fig. 7*B*). We observe that the peptides have a moderate to very high spatial correlation in both the clusters (Fig. 7*B*), thereby providing an additional and strong validation of the findings of ImShot.

ImShot: the Desktop Application and GUI

ImShot GUI has two parts: sidebar and main panels (Fig. 8*A*). The sidebar contains the different modules of the software as dropdown menus. For the two-group comparison (Limma), it renders high-resolution interactive volcano plot along with numeric input boxes for modifying fold change and false discovery rate that allow users for desired data thresholding (Fig. 8*B*). In addition, the user has the option of not adjusting *p* values when false negatives are very costly (See "Statistical Analysis of Proteomics Datasets Using Moderated *t* Test" section in Computational Methods section; Fig. 8*B*). Users can also save the plot in the PNG file format. ImShot also shows that the Limma-moderated *t* test results in the form of a searchable table, which can be exported as excel or csv format (Fig. 8*C*).

For the GO and pathway enrichment analyses, it creates high-resolution plots of GO–gene and pathway–pathway interaction networks using the top 10 most significant GO and pathway terms, respectively (Fig. 8, *D* and *F*). The plots can be exported as PNG image files. These plots in the GUI are zoomable, and the nodes are highly flexible allowing the users to select nodes of their choice for rearranging them freely (see video tutorials and supplementary file) to create a network map according to their convenience and need. ImShot also shows the over-representation test results in the form of a searchable table (for top 10 most significant GO terms/pathways) below the network plot, which can be exported as excel or csv format (Fig. 8, *E* and *G*). To provide additional flexibility when plotting the results of GO and pathway analyses, we included a function in the ImShot R package that takes the resulting data structure (after GO/pathway analysis) and plots the graphs directly on a Cytoscape (52) session using the user-specified font family and size.



Copy CSV Excel Show 10 entries

Table: Limma t-test results Search:

Uniprot	Gene_name	logFC	p_mod
A1EGX6	Fscb	-2-2179	0.0212
A2A432	Cul4b	-0.2469	0.4813
A2A6C4	Sppl2c	-1.7605	0.2052
A2A8Z1	Osbpl9	0.1778	0.7279

Copy CSV Excel Show 10 entries

Table: GO enrichment analysis results Search:

ID	Description	GeneRatio	BgRatio	pvalue
GO:0006457	protein folding	4/24	145/15416	0.000068
GO:0006843	mitochondrial citrate transmembrane transport	1/24	1/15416	0.001556
GO:0006986	response to unfolded protein	3/24	10/15416	0.000591
GO:0007283	spermatogenesis	5/24	485/15416	0.000781

Copy CSV Excel Show 10 entries

Table: Pathway enrichment analysis results Search:

ID	Description	GeneRatio	BgRatio	pvalue
R-MMU-142851	TCA cycle and electron transport	4/24	138/7819	0.00075
R-MMU-1442490	Collagen degradation	3/24	45/7819	0.00033
R-MMU-1474290	Collagen formation	3/24	79/7819	0.00017
R-MMU-1592230	Mitochondrial biogenesis	2/24	28/7819	0.00032

FIG. 8. **ImShot GUI.** *A*, ImShot GUI sidebar (*yellow dashed rectangle*) and main panel (*red dashed rectangle*). The *sidebar panel* can toggle upon clicking on the icon enclosed in *white dashed rectangle*. *B*, interactive volcano plot. *Reddish dots* indicate statistically significant proteins after Limma-moderated *t* test (($lfc > 2 \parallel lfc < -2$)) and $p < 0.05$ and *q* value-based FDR control. The plot updates automatically when the lfc/FDR is tuned using the input boxes provided. In addition, the plot can be updated according to whether a *p* value adjustment is desired. Placing the cursor on a data point of the volcano plot provides information about the protein identity, its fold change, and *p* value (*reddish rectangle*). *C*, searchable table after Limma-moderated statistics. *D*, protein–GO term interaction network after over-representation test. *Tiny dark colored fixed sized nodes* represent proteins, and *light colored variable sized node*s represent GO terms, and their sizes are proportional to the numbers of proteins involved in them. A protein is connected to a GO term *via* an edge if and only if the term is enriched (*p* value is adjusted) in that protein. *E*, searchable table after GO analysis. *F*, pathway–pathway interaction network after over-representation test. Two pathways are connected *via* an edge if and only if both are enriched (*p* value is adjusted) in at least one common protein. Size of a node is proportional to the number of proteins involved in it, and the color is proportional to the adjusted *p* value; lower *p* value maps to darker color. *G*, searchable table after pathway over-representation test. Contents from all the tables can be copied into the clipboard or exported as CSV/EXCEL by clicking corresponding buttons on top of the tables (shown in *dashed red rectangle*). FDR, false discovery rate; GO, Gene Ontology; GUI, graphic user interface; lfc, log fold change.

In addition, ImShot maintains an operation log (supplemental Fig. S3) that allow users to record all the steps along with names of the files and version of R used. This paves the way for the user to reproduce the data analysis later without any ambiguity.

DISCUSSION

ImShot is the first software of its kind to provide an end-to-end analysis by integrating two orthogonal MS technologies. The software can be used in any two-group comparison test, that is, animal models, patient samples, and so on, allowing user flexibility in terms of experimental context. The software deals with both the IMS and LC–MS data and integrates them through a GUI that does not require any computational expertise to operate. Although there have been studies aimed at identifying peptides from IMS datasets, the approaches have been diverse, and not all of them involved a combination with LC–MS. Most of these approaches involved a considerable manual intervention resulting in a very limited output. The others relied either on customized experimental approaches or specific IMS platforms. Moreover, almost all the studies lacked a defined and conveniently executable strategy that could be used for integrating the two complementary datasets in general, especially in the diseased context (24–32, 59) (supplemental Table S2). Implementation of ImShot is independent of IMS platforms used to generate the data. The simple requirement of a mass list that characterizes a spatial peptide cluster and a ranking algorithm that accounts for the incompatibility in mass resolution between IMS and LC–MS measurements gives the much desired generality in its use. In addition, searching IMS masses in the corresponding LC–MS dataset is associated with a user-defined tolerance, which gives the user flexibility to use ImShot on a wide variety of data generated by different IMS platforms. As a result, ImShot can be used not only as an efficient high-throughput screening tool but can also be integrated in experimental strategies targeted at robust *in situ* peptide identifications (29, 31, 59).

While dealing with IMS data, ImShot performs a very crucial task of deisotoping the peptide spectra based on spatial data segregation. In absence of deisotoping, the resulting IMS spectra would be biased toward an overestimation of the number of peptide peaks and will also include ambiguous annotations of peptide masses when comparing with LC–MS data. To the best of our knowledge, this is the only software that deisotopes IMS peptide spectra to get rid of false positives. The novel method of ranking of IMS peptides in case of multiple annotations (based on our proposed MLP scoring) associates most likely biological pathways with the most probable areas of the tissue. Furthermore, because ImShot only considers peptides that are commonly found in the two different ionization procedures (MALDI and ESI in this example), the results are most likely unaffected by the apparent disparity in peptide intensities introduced by the separate ionizations (60–62). ImShot also employs a combination of three discriminatory qualities of a peptide (mean intensity, log fold change, and *p* value) for ranking and final identification, reducing the importance of exclusive technical aspects. Computational, experimental, literature, and database-based validation (Figs. 6 and 7) of the ranking method has imparted sufficient confidence in our scoring approach, which can now be applied to any type of tissues for two-group comparison test.

ImShot provides users with a wide choice of data processing options when it comes to LC–MS data. This not only allows the LC–MS data to be compared with the related IMS data but also provides the option of performing a comprehensive LC–MS data analysis separately. The algorithm design minimizes the contribution of missing values in the final data and at the same time provides a multitude of possibilities for imputing those values. Many GUI-based software for LC–MS data analysis have either been desktop applications running only on Windows platform (35, 63, 64) or web applications (65–67). Though web applications have many benefits, their stability depends on the state of the server running the application, number of users accessing it, network bandwidth, and so on. The web applications are often written using shiny R package (65, 66) that are not easy to debug https://shiny.rstudio.com/articles/debugging.html; (Accessed on 2021/07/05). The general application of breakpoints for debugging is largely unsuccessful for shiny applications as it has very restricted usage. Alternatively, one can call the *browser* function that interrupts code execution and facilitates investigation of the programming environment. However, it involves manual removal of the *browser* function for uninterrupted execution of the software, thereby burdening the developer with an additional housekeeping step. On the other hand, the desktop applications so far have been lacking the aesthetics in generating graphical output leading users to write additional scripts or use graphics editing software to make publication quality figures. In ImShot, we have tried to incorporate the best of the two worlds in producing high-quality graphics like a web application and at the same time running natively on user's computer.

ImShot works like a native web application that can read and write data besides accessing the computer’s file system. Moreover, Electron framework saves time by providing a large pool of APIs, which the developers can utilize in their desktop applications conveniently. ImShot is an open-source software under Massachusetts Institute of Technology license that allows anyone to view and modify its source codes to adapt or extend it to use in more customized environments. The modular design of the front ends and back ends of ImShot allows time-efficient implementation of new features. Other major advantages of this mode include code optimization, modularity, faster deployment, and flexibility in switching programming languages in the back end or changing frameworks in the front end. Modular architecture of ImShot’s codebase enables each function to perform a specific task, thereby allowing the modules to be used independently. Since ImShot performs lots of statistical computations in the back end, the use of R makes a perfect choice, and usage of HTML, CSS, and JS in the front end makes the software extremely flexible and feature rich. In addition, it also records the R code runtime, which allows a software developer to monitor and optimize (if needed) the back end. ImShot desktop application provides question mark icons with hover effects next to every interface element (input, dropdown box, file upload wizard, etc) to guide users through the meaning of the input.

In addition, we have released ImShot on GitHub, where users can request feature(s) for the software and/or report issues. This is part of our plans to augment many more functionalities in the future releases of ImShot and make it most useful and convenient for the research community.

DATA AVAILABILITY

The ImShot desktop application, R package, test datasets, and video tutorials along with software documentation are available in this GitHub repository: https://github.com/wasimaftab/ImShot. See supplementary file for the instructions on how to set up and run the software.

Supplemental data—This article contains supplemental data (69–76).

Acknowledgments—We thank Prof Matti Poutanen from the Institute of Biomedicine, University of Turku, Finland and Prof Dr Artur Mayerhofer from the Cell Biology-Anatomy III of Biomedical Center, LMU Munich, Germany for making available the mouse model(s) and the tissues used as the experimental basis to build ImShot. We thank Marco Borso, Anuroop Venkateswaran, Klara Krmpotic, and Frederike Schäfer from the Imhof group for their critical feedback during developing ImShot. Thanks are due to Dr Parijat Majumder of the Max-Planck Institute of Biochemistry, Martinsried, Germany for her help during preparation of the figures presented in this article.

Author contributions—S. L. and A. I. conceptualization; W. A. and S. L. methodology; W. A. and S. L. software; W. A. and S. L. validation; S. L. investigation; A. I. resources; W. A., S. L., and A. I. writing–original draft; W. A., S. L., and A. I. writing–review & editing; W. A. and S. L. visualization; S. L. and A. I. supervision; A. I. funding acquisition.

Funding and additional information—This work was supported by grants from the Deutsche Forschungsgemeinschaft (grant nos.: CRC1123-TPZ02, CRC1309-TPB03, and CRC1064-TPZ03; to A. I.) and predoctoral grants (QBM; to W. A.).

Conflict of interest—The authors declare no competing interests.

Abbreviations—The abbreviations used are: API, application programming interface; BP, biological process; CSS, cascading style sheet; ECM, extracellular matrix; GO, Gene Ontology; GUI, graphic user interface; HTML, hypertext markup language; IMS, imaging mass spectrometry; JS, JavaScript; JSON, JS Object Notation; MLP, most likely peptide; OS, operating system.

Received September 10, 2021, and in revised from, April 8, 2022 Published, MCPRO Papers in Press, May 13, 2022, https://doi.org/10.1016/j.mcpro.2022.100242

REFERENCES

1. Key, J., Mueller, A. K., Gispert, S., Matschke, L., Wittig, I., Corti, O., Münch, C., Decher, N., and Auburger, G. (2019) Ubiquitylome profiling of Parkin-null brain reveals dysregulation of calcium homeostasis factors ATP1A2, Hippocalcin and GNA11, reflected by altered firing of noradrenergic neurons. *Neurobiol. Dis.* **127**, 114–130
2. Pineau, C., Hikmet, F., Zhang, C., Oksvold, P., Chen, S., Fagerberg, L., Uhlén, M., and Lindskog, C. (2019) Cell type-specific expression of testis elevated genes based on transcriptomics and antibody-based proteomics. *J. Proteome Res.* **18**, 4215–4230
3. Fawkner-Corbett, D., Antanaviciute, A., Parikh, K., Jagielowicz, M., Gerós, A. S., Gupta, T., Ashley, N., Khamis, D., Fowler, D., and Morrissey, E. (2021) Spatiotemporal analysis of human intestinal development at single-cell resolution. *Cell* **184**, 810–826.e823
4. Petyuk, V. A., Qian, W.-J., Chin, M. H., Wang, H., Livesay, E. A., Monroe, M. E., Adkins, J. N., Jaitly, N., Anderson, D. J., and Camp, D. G. (2007) Spatial mapping of protein abundances in the mouse brain by voxelation integrated with high-throughput liquid chromatography–mass spectrometry. *Genome Res.* **17**, 328–336
5. Calabresi, P., Centonze, D., Pisani, A., and Bernardi, G. (1999) Metabotropic glutamate receptors and cell-type-specific vulnerability in the striatum: Implication for ischemia and huntington's disease. *Exp. Neurol.* **158**, 97–108
6. Lin, Y.-T., Seo, J., Gao, F., Feldman, H. M., Wen, H.-L., Penney, J., Cam, H. P., Gjoneska, E., Raja, W. K., and Cheng, J. (2018) APOE4 causes widespread molecular and cellular alterations associated with Alzheimer's disease phenotypes in human iPSC-derived brain cell types. *Neuron* **98**, 1141–1154.e7
7. Lahiri, S., Aftab, W., Walenta, L., Strauss, L., Poutanen, M., Mayerhofer, A., and Imhof, A. (2021) MALDI-IMS combined with shotgun proteomics identify and localize new factors in male infertility. *Life Sci. Alliance* **4**, e202000672
8. Parker, N. R., Khong, P., Parkinson, J. F., Howell, V. M., and Wheeler, H. R. (2015) Molecular heterogeneity in glioblastoma: Potential clinical implications. *Front. Oncol.* **5**, 55
9. Pinato, D. J., Shiner, R. J., White, S. D. T., Black, J. R. M., Trivedi, P., Stebbing, J., Sharma, R., and Mauri, F. A. (2016) Intra-tumoral heterogeneity in the expression of programmed-death (PD) ligands in isogeneic primary and metastatic lung cancer: Implications for immunotherapy. *Oncoimmunology* **5**, e1213934
10. Evrard, C., Tachon, G., Randrian, V., Karayan-Tapon, L., and Tougeron, D. (2019) Microsatellite instability: Diagnosis, heterogeneity, discordance, and clinical impact in colorectal cancer. *Cancers* **11**, 1567
11. y Cajal, S. R., Sesé, M., Capdevila, C., Aasen, T., De Mattos-Arruda, L., Diaz-Cano, S. J., Hernández-Losa, J., and Castellví, J. (2020) Clinical implications of intratumor heterogeneity: Challenges and opportunities. *J. Mol. Med.* **98**, 161–177
12. Stoeckli, M., Chaurand, P., Hallahan, D. E., and Caprioli, R. M. (2001) Imaging mass spectrometry: A new technology for the analysis of protein expression in mammalian tissues. *Nat. Med.* **7**, 493–496
13. Ishikawa, S., Tateya, I., Hayasaka, T., Masaki, N., Takizawa, Y., Ohno, S., Kojima, T., Kitani, Y., Kitamura, M., and Hirano, S. (2012) Increased expression of phosphatidylcholine (16: 0/18: 1) and (16: 0/18: 2) in thyroid papillary cancer. *PLoS one* **7**, e48873
14. Chaurand, P., Norris, J. L., Cornett, D. S., Mobley, J. A., and Caprioli, R. M. (2006) New developments in profiling and imaging of proteins from tissue sections by MALDI mass spectrometry. *J. Proteome Res.* **5**, 2889–2900
15. Jones, E. E., Powers, T. W., Neely, B. A., Cazares, L. H., Troyer, D. A., Parker, A. S., and Drake, R. R. (2014) MALDI imaging mass spectrometry profiling of proteins and lipids in clear cell renal cell carcinoma. *Proteomics* **14**, 924–935
16. Meistermann, H., Norris, J. L., Aerni, H.-R., Cornett, D. S., Friedlein, A., Erskine, A. R., Augustin, A., Mudry, M. C. D. V., Ruepp, S., and Suter, L. (2006) Biomarker discovery by imaging mass spectrometry: Transthyretin is a biomarker for gentamicin-induced nephrotoxicity in rat. *Mol. Cell Proteomics* **5**, 1876–1886

17. Rauser, S., Marquardt, C., Balluff, B., Deininger, S.-O., Albers, C., Belau, E., Hartmer, R., Suckau, D., Specht, K., and Ebert, M. P. (2010) Classification of HER2 receptor status in breast cancer tissues by MALDI imaging mass spectrometry. *J. Proteome Res.* **9**, 1854–1863
18. Gustafsson, J. O. R., Oehler, M. K., Ruszkiewicz, A., McColl, S. R., and Hoffmann, P. (2011) MALDI imaging mass spectrometry (MALDI-IMS)—application of spatial proteomics for ovarian cancer classification and diagnosis. *Int. J. Mol. Sci.* **12**, 773–794
19. Balluff, B., Hanselmann, M., and Heeren, R. M. A. (2017) Mass spectrometry imaging for the investigation of intratumor heterogeneity. *Adv. Cancer Res.* **134**, 201–230
20. de San Roman, E. G., Manuel, I., Giralt, M. T., Ferrer, I., and Rodríguez-Puertas, R. (2017) Imaging mass spectrometry (IMS) of cortical lipids from preclinical to severe stages of Alzheimer's disease. *Biochim. Biophys. Acta (BBA)-Biomem.* **1859**, 1604–1614
21. Pauker, V. I., Bertzbach, L. D., Hohmann, A., Kheimar, A., Teifke, J. P., Mettenleiter, T. C., Karger, A., and Kaufer, B. B. (2019) Imaging mass spectrometry and proteome analysis of Marek's disease virus-induced tumors. *Msphere* **4**, e00569-18
22. Palmer, A., Phapale, P., Chernyavsky, I., Lavigne, R., Fay, D., Tarasov, A., Kovalev, V., Fuchser, J., Nikolenko, S., and Pineau, C. (2017) FDR-controlled metabolite annotation for high-resolution imaging mass spectrometry. *Nat. Met.* **14**, 57–60
23. Alberts, D., Pottier, C., Smargiasso, N., Baiwir, D., Mazzucchelli, G., Delvenne, P., Kriegsmann, M., Kazdal, D., Warth, A., and De Pauw, E. (2018) MALDI imaging-guided microproteomic analyses of heterogeneous breast tumors—a pilot study. *Proteomics Clin. Appl.* **12**, 1700062
24. Longuespée, R., Ly, A., Casadonte, R., Schwamborn, K., Kazdal, D., Zgorzelski, C., Bollwein, C., Kriegsmann, K., Weichert, W., and Kriegsmann, J. (2019) Identification of MALDI imaging proteolytic peptides using LC-MS/MS-based biomarker discovery data: A proof of concept. *Proteomics Clin. Appl.* **13**, 1800158
25. Schober, Y., Guenther, S., Spengler, B., and Römpp, A. (2012) High-resolution matrix-assisted laser desorption/ionization imaging of tryptic peptides from tissue. *Rapid Commun. Mass Spectrom.* **26**, 1141–1146
26. Huber, K., Khamehgir-Silz, P., Schramm, T., Gorshkov, V., Spengler, B., and Römpp, A. (2018) Approaching cellular resolution and reliable identification in mass spectrometry imaging of tryptic peptides. *Anal. Bioanal. Chem.* **410**, 5825–5837
27. Groseclose, M. R., Andersson, M., Hardesty, W. M., and Caprioli, R. M. (2007) Identification of proteins directly from tissue: *In situ* tryptic digestions coupled with imaging mass spectrometry. *J. Mass Spectrom.* **42**, 254–262
28. Franck, J., El Ayed, M., Wisztorski, M., Salzet, M., and Fournier, I. (2009) On-tissue N-terminal peptide derivatizations for enhancing protein identification in MALDI mass spectrometric imaging strategies. *Anal. Chem.* **81**, 8305–8317
29. Heijs, B., Carreira, R. J., Tolner, E. A., de Ru, A. H., van den Maagdenberg, A. M., van Veelen, P. A., and McDonnell, L. A. (2015) Comprehensive analysis of the mouse brain proteome sampled in mass spectrometry imaging. *Anal. Chem.* **87**, 1867–1875
30. Maier, S. K., Hahne, H., Gholami, A. M., Balluff, B., Meding, S., Schoene, C., Walch, A. K., and Kuster, B. (2013) Comprehensive identification of proteins from MALDI imaging. *Mol. Cell Proteomics* **12**, 2901–2910
31. Piehowski, P. D., Zhu, Y., Bramer, L. M., Stratton, K. G., Zhao, R., Orton, D. J., Moore, R. J., Yuan, J., Mitchell, H. D., and Gao, Y. (2020) Automated mass spectrometry imaging of over 2000 proteins from tissue sections at 100-μm spatial resolution. *Nat. Commun.* **11**, 1–12
32. Guo, G., Papanicolaou, M., Demarais, N., Wang, Z., Schey, K., Timpson, P., Cox, T., and Grey, A. (2021) Automated annotation and visualisation of high-resolution spatial proteomic mass spectrometry imaging data using HIT-MAP. *Nat. Commun.* **12**, 1–16
33. Trede, D., Schiffler, S., Becker, M., Wirtz, S., Steinhorst, K., Strehlow, J., Aichler, M., Kobarg, J. H., Oetjen, J., and Dyatlov, A. (2012) Exploring three-dimensional matrix-assisted laser desorption/ionization imaging mass spectrometry data: Three-dimensional spatial segmentation of mouse kidney. *Anal. Chem.* **84**, 6079–6087
34. Tyanova, S., Temu, T., and Cox, J. (2016) The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat. Protoc.* **11**, 2301–2319
35. Tyanova, S., Temu, T., Sinitcyn, P., Carlson, A., Hein, M. Y., Geiger, T., Mann, M., and Cox, J. (2016) The Perseus computational platform for comprehensive analysis of (prote) omics data. *Nat. Met.* **13**, 731–740
36. [preprint] Ahlmann-Eltze, C., and Anders, S. (2020) proDA: Probabilistic dropout analysis for identifying differentially abundant proteins in label-free mass spectrometry. *biorxiv*. https://doi.org/10.1101/661496
37. Kammers, K., Cole, R. N., Tiengwe, C., and Ruczinski, I. (2015) Detecting significant changes in protein abundance. *EuPA Open Proteomics* **7**, 11–19
38. Lönnstedt, I., and Speed, T. (2002) Replicated microarray data. *Stat. Sinica* **12**, 31–46
39. Smyth, G. K. (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* **3**. https://doi.org/10.2202/1544-6115.1027
40. Yu, L., Gulati, P., Fernandez, S., Pennell, M., Kirschner, L., and Jarjoura, D. (2011) Fully moderated T-statistic for small sample size gene expression arrays. *Stat. Appl. Genet. Mol. Biol.* **10**, 42
41. Brusniak, M.-Y., Bodenmiller, B., Campbell, D., Cooke, K., Eddes, J., Garbutt, A., Lau, H., Letarte, S., Mueller, L. N., and Sharma, V. (2008) Corra: Computational framework and tools for LC-MS discovery and targeted mass spectrometry-based proteomics. *BMC Bioinform.* **9**, 542
42. Salvatori, R., Aftab, W., Forne, I., Imhof, A., Ott, M., and Singh, A. P. (2020) Mapping protein networks in yeast mitochondria using proximity-dependent biotin identification coupled to proteomics. *STAR Protoc.* **1**, 100219
43. Schwammle, V., León, I. R., and Jensen, O. N. (2013) Assessment and improvement of statistical tools for comparative proteomics analysis of sparse data sets with few experimental replicates. *J. Proteome Res.* **12**, 3874–3883
44. Ting, L., Cowley, M. J., Hoon, S. L., Guilhaus, M., Raftery, M. J., and Cavicchioli, R. (2009) Normalization and statistical analysis of quantitative proteomics data generated by metabolic labeling. *Mol. Cell Proteomics* **8**, 2227–2242
45. van Ooijen, M. P., Jong, V. L., Eijkemans, M. J., Heck, A. J., Andeweg, A. C., Binai, N. A., and van den Ham, H.-J. (2018) Identification of differentially expressed peptides in high-throughput proteomics data. *Brief. Bioinform.* **19**, 971–981
46. Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., and Smyth, G. K. (2015) Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucl. Acids Res.* **43**, e47
47. Storey, J. D. (2003) The positive false discovery rate: A bayesian interpretation and the q-value. *Ann. Stat.* **31**, 2013–2035
48. Storey, J. D., and Tibshirani, R. (2003) Statistical significance for genome-wide studies. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 9440–9445
49. McDonald, J. H. (2009) *Handbook of Biological Statistics*. sparky house publishing Baltimore, MD
50. Strohalm, M., Kavan, D., Novak, P., Volny, M., and Havlicek, V. (2010) mMass 3: A cross-platform software environment for precise analysis of mass spectrometric data. *Anal. Chem.* **82**, 4648–4651
51. Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012) clusterProfiler: An R package for comparing biological themes among gene clusters. *OMICS.* **16**, 284–287
52. Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003) Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504
53. Yu, G., and He, Q.-Y. (2016) ReactomePA: An R/bioconductor package for reactome pathway analysis and visualization. *Mol. BioSyst.* **12**, 477–479
54. Adam, M., Urbanski, H. F., Garyfallou, V. T., Welsch, U., Köhn, F. M., Ullrich Schwarzer, J., Strauss, L., Poutanen, M., and Mayerhofer, A. (2012) High levels of the extracellular matrix proteoglycan decorin are associated with inhibition of testicular function. *Int. J. Androl.* **35**, 550–561
55. Alfano, M., Pederzoli, F., Locatelli, I., Ippolito, S., Longhi, E., Zerbi, P., Ferrari, M., Brendolan, A., Montorsi, F., and Drago, D. (2019) Impaired testicular signaling of vitamin A and vitamin K contributes to the aberrant composition of the extracellular matrix in idiopathic germ cell aplasia. *Fertil. Ster.* **111**, 687–698
56. Boyle, E. I., Weng, S., Gollub, J., Jin, H., Botstein, D., Cherry, J. M., and Sherlock, G. (2004) GO:: TermFinder—open source software for accessing gene ontology information and finding significantly enriched

gene ontology terms associated with a list of genes. *Bioinformatics* **20**, 3710–3715

57. Mayer, C., Adam, M., Glashauser, L., Dietrich, K., Schwarzer, J. U., Köhn, F. M., Strauss, L., Welter, H., Poutanen, M., and Mayerhofer, A. (2016) Sterile inflammation as a factor in human male infertility: Involvement of Toll like receptor 2, biglycan and peritubular cells. *Sci. Rep.* **6**, 1–10
58. Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., and Asplund, A. (2015) Tissue-based map of the human proteome. *Science* **347**, 1260419
59. Meding, S., Martin, K., Gustafsson, O. J., Eddes, J. S., Hack, S., Oehler, M. K., and Hoffmann, P. (2013) Tryptic peptide reference data sets for MALDI imaging mass spectrometry on formalin-fixed ovarian cancer tissues. *J. Proteome Res.* **12**, 308–315
60. Nadler, W. M., Waidelich, D., Kerner, A., Hanke, S., Berg, R., Trumpp, A., and Rösli, C. (2017) MALDI versus ESI: The impact of the ion source on peptide identification. *J. Proteome Res.* **16**, 1207–1215
61. Person, M. D., Lo, H.-H., Towndrow, K. M., Jia, Z., Monks, T. J., and Lau, S. S. (2003) Comparative identification of prostanoid inducible proteins by LC-ESI-MS/MS and MALDI-TOF mass spectrometry. *Chem. Res. Toxicol.* **16**, 757–767
62. Lim, H., Eng, J., Yates, J. R., Tollaksen, S. L., Giometti, C. S., Holden, J. F., Adams, M. W., Reich, C. I., Olsen, G. J., and Hays, L. G. (2003) Identification of 2D-gel proteins: A comparison of MALDI/TOF peptide mass mapping to μ LC-ESI tandem mass spectrometry. *J. Am. Soc. Mass Spectrom.* **14**, 957–970
63. Rigbolt, K. T., Vanselow, J. T., and Blagoev, B. (2011) GProX, a user-friendly platform for bioinformatics analysis and visualization of quantitative proteomics data. *Mol. Cell Proteomics* **10**. https://doi.org/10.1074/mcp.O110.007450
64. Chang, C., Xu, K., Guo, C., Wang, J., Yan, Q., Zhang, J., He, F., and Zhu, Y. (2018) PANDA-view: An easy-to-use tool for statistical analysis and visualization of quantitative proteomics data. *Bioinformatics* **34**, 3594–3596
65. Weiner, A. K., Sidoli, S., Diskin, S. J., and Garcia, B. A. (2018) Graphical interpretation and analysis of proteins and their ontologies (GiaPronto): A one-click graph visualization software for proteomics data sets. *Mol. Cell Proteomics* **17**, 1426–1431
66. Gallant, J. L., Heunis, T., Sampson, S. L., and Bitter, W. (2020) ProVision: A web-based platform for rapid analysis of proteomics data processed by MaxQuant. *Bioinformatics* **36**, 4965–4967
67. Efstathiou, G., Antonakis, A. N., Pavlopoulos, G. A., Theodosiou, T., Divanach, P., Trudgian, D. C., Thomas, B., Papanikolaou, N., Aivaliotis, M., and Acuto, O. (2017) ProteoSign: An end-user online differential proteomics statistical analysis platform. *Nucl. Acids Res.* **45**, W300–W306
68. Bult, C. J., Blake, J. A., Smith, C. L., Kadin, J. A., and Richardson, J. E. (2019) Mouse genome database (MGD) 2019. *Nucl. Acids Res.* **47**, D801–D806
69. Verboven, S., Branden, K. V., and Goos, P. (2007) Sequential imputation for missing values. *Comput. Biol. Chem.* **31**, 320–327
70. Branden, K. V., and Verboven, S. (2009) Robust data imputation. *Comput. Biol. Chem.* **33**, 7–13
71. Stekhoven, D. J., and Bühlmann, P. (2012) MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics* **28**, 112–118
72. Josse, J., and Husson, F. (2012) Handling missing values in exploratory multivariate data analysis methods. *Journal de la Société Française de Statistique* **153**, 79–99
73. Roweis, S. (1997) EM algorithms for PCA and SPCA. *Adv. Neural Inf. Process Syst.* **10**, 626–632
74. Josse, J., Pagès, J., and Husson, F. (2011) Multiple imputation in principal component analysis. *Adv. Data Anal. Classif.* **5**, 231–246
75. Oba, S., Sato, M.-A., Takemasa, I., Monden, M., Matsubara, K.-I., and Ishii, S. (2003) A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics* **19**, 2088–2096
76. Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. B. (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics* **17**, 520–525