UC San Diego

UC San Diego Previously Published Works

Title

ProteoStorm: An Ultrafast Metaproteomics Database Search Framework.

Permalink

https://escholarship.org/uc/item/0kr5f7xs

Journal

Cell Systems, 7(4)

ISSN

2405-4712

Authors

Beyter, Doruk

Lin, Miin

Yu, Yanbao

et al.

Publication Date

2018-10-24

DOI

10.1016/j.cels.2018.08.009

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

Cell Syst. Author manuscript; available in PMC 2019 October 24.

Published in final edited form as:

Cell Syst. 2018 October 24; 7(4): 463–467.e6. doi:10.1016/j.cels.2018.08.009.

ProteoStorm: An ultrafast metaproteomics database search framework

Doruk Beyter[#1,4], **Miin S. Lin**[#2], **Yanbao Yu**[3], **Rembert Pieper**[3], and **Vineet Bafna**[1,5,*]

[1]Department of Computer Science and Engineering, University of California San Diego, La Jolla, California, 92093, USA [2]Graduate Program in Bioinformatics & Systems Biology, University of California San Diego, La Jolla, California, 92093, USA [3]The J. Craig Venter Institute, Rockville, Maryland, 20850, USA [4]Current Address: deCODE Genetics/Amgen, Inc., Reykjavik, Iceland [5]Lead Contact

[#] These authors contributed equally to this work.

Summary

Shotgun metaproteomics has potential to reveal the functional landscape of microbial communities, but lacks appropriate methods for complex samples with unknown compositions. In the absence of prior taxonomic information, tandem mass spectra would be searched against large pan-microbial databases, which requires heavy computational workload and reduces sensitivity. We present ProteoStorm, an efficient database search framework for large-scale metaproteomics studies, which identifies high-confidence peptide-spectrum matches (PSMs) while achieving a two to three orders-of-magnitude speedup over popular tools. A reanalysis of a urinary tract infection (UTI) dataset of 110 individuals revealed a complex pattern of polymicrobial expression, including sub-types of urinary tract infections, cases of bacterial vaginosis, and evidence of no underlying disease. Importantly, compared to the initial UTI study that restricted the search database to a manually-curated list of 20 genera, ProteoStorm identified additional genera that were previously unreported, including a case of infection with the rare pathogen *Propionimicrobium*.

eTOC Blurb:

[*]Correspondence: vbafna@cs.ucsd.edu.

Author Contributions

D.B., M.S.L., and V.B. conceived, designed, and developed the ProteoStorm database search framework. Specifically, D.B. developed the algorithmic strategy (ion-mass indexing), and M.S.L. developed the software strategy (spectral and database partitioning) for peptide filtering with guidance from V.B.. D.B. and M.S.L. enabled the *p*-value computation between given peptide-spectrum pairs via modifying MS-GF+. M.S.L. wrote the command-line interface for ProteoStorm, and produced results. M.S.L., Y.Y., R.P. and V.B. provided biological interpretation of the UTI data. D.B., M.S.L., R.P., and V.B. wrote the manuscript with feedback from all authors.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

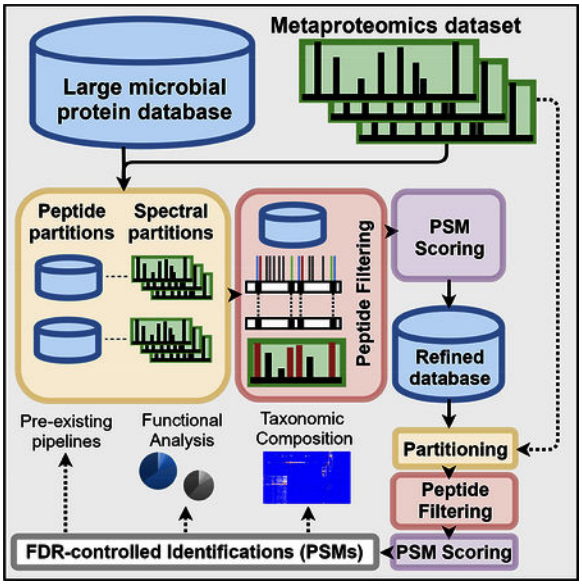
Declaration of Interests

V.B. is a cofounder, has an equity interest in, and receives income from Digital Proteomics, LLC. The terms of this arrangement have been reviewed and approved by the University of California, San Diego, in accordance with its conflict-of-interest policies. Digital Proteomics was not involved in the research presented here.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

In the absence of prior taxonomic information, tandem mass spectra would be searched against large pan-microbial databases, requiring heavy computational workload. We present ProteoStorm, an efficient database search framework for large-scale metaproteomics studies, achieving a two to three orders-of-magnitude speedup over popular tools, while maintaining high sensitivity. A reanalysis of a urinary tract infection (UTI) dataset revealed complex pattern of polymicrobial expression and previously unreported rare pathogens.

Graphical Abstract



Keywords

proteomics; metaproteomics; microbiology; microbial communities; LC-MS/MS; proteome informatics

Metaproteomics is an emerging field that identifies expressed proteins using tandem mass spectrometry (MS/MS) to decipher the taxonomic and functional realm of complex microbial environments. Without prior knowledge of the active organisms in a sample, downstream analyses are highly dependent on the accurate assignment of MS/MS spectra to peptide sequences from massive (~10Gb+) pan-microbial protein databases. While iterative search methods utilizing conventional database search tools (Jagtap, et al., 2013; Tang, et al., 2016; Zhang, et al., 2016) increase identifications by reducing the initial search space, and methods such as ComPIL (Chatterjee, et al., 2016) address heavy computational workload by distributing data across multiple servers, searching large databases in reasonable time frames, without the requirement of high computational capacity, remains a major impediment. We present ProteoStorm, an ultrafast metaproteomics database search framework that utilizes a multi-staged, efficient filtering of peptide-spectrum matches (PSMs). On large pan-microbial databases and spectral datasets, ProteoStorm achieved a speedup by two to three orders-of-magnitude over popular database search tools while maintaining high sensitivity in terms of peptides identified.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

There are many reasons why conventional database search tools do not scale well in metaproteomics studies. For a large database, D, a common practice is to split D into k arbitrary small files d_i, where each of m spectral files is searched against each d_i. While memory efficient, this practice results in the inefficient and redundant search of spectra against peptides that would never result in a viable match. ProteoStorm assumes that for each detectable protein in a microbial sample, there exists at least one fully-tryptic peptide with no variable modification, identifiable with a sufficiently low p-value PSM.

This motivates a multi-stage strategy (Fig. 1a), where a fast, fully-tryptic search in the first stage is followed by a semi-tryptic search in the second stage that is limited to proteins identified in the first stage. Each stage of ProteoStorm is composed of three core modules: i) database and spectra partitioning, ii) efficient and sensitive peptide filtering, and iii) PSM p-value computation (STAR Methods).

In the first module, ProteoStorm bins the unique set of *in silico* digested tryptic peptides from a pan-microbial database, D, into n database partitions with pre-defined mass windows in Daltons (Fig. 1b; STAR Methods). Separately, spectra are organized into n matching partitions, based on their parent masses. Each spectral partition is searched only against its matching database partition (instead of k database files mentioned above), reducing the number of computations dramatically. Moreover, the partitioning of the entire pan-microbial database is a one-time operation, and the partitions can be used for any metaproteomics experiment.

To overcome the inefficiency of scoring spectra against peptides with poor match of b-/y-ions, the second module of ProteoStorm filters spectrum-peptide pairs with insufficient number of shared peaks using an ion-mass indexing data structure, I (Fig. 1c; STAR Methods). Similar to previous studies (Rinner, et al., 2008; Stein, 1995; Burke, et al., 2017; Kong, et al., 2017), ProteoStorm utilizes the fact that theoretical ions from different peptides may share the same m/z value given a charge and fragment mass tolerance. For each database partition, ProteoStorm constructs I, where ion-mass index i references a mass-sorted list of peptides containing a b-/y-ion within the mass tolerance of i. For each prominent peak in a spectrum s, its corresponding ion-mass index $i \in I$ is accessed to retrieve the collection of peptides, P, containing a matching ion. For all peptides $p \in P$ that are within the parent mass tolerance of the spectrum, the shared peak count (SPC) of the spectrum-peptide pair (s,p) is incremented by one. Spectrum-peptide pairs with a sufficient SPC are retained for further analysis in the third module. This ion-mass indexing-based peptide filtering enables ultra-fast querying of a spectrum against a peptide database as it bypasses all peptide ions with no matching spectrum peak, and matches prominent spectral peaks against all peptide ions simultaneously.

In the third module, ProteoStorm utilizes the MS-GF+ generating function (Kim & Pevzner, 2014) to compute p-values for (s,p) pairs with the highest MS-GF+ raw score for a given spectrum s. Accurate estimation of p-values allows for the subsequent computation of a peptide-level FDR. Peptides identified in the first stage are used to construct a refined protein database for a semi-tryptic second-stage search (STAR Methods).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

We evaluated the performance of ProteoStorm using 7.5 million LC-MS/MS spectra obtained from urinary pellets (UP) of 110 suspected urinary tract infection (UTI) cases and five healthy individuals from two related studies (Yu, et al., 2015; Yu, et al., 2017). When searching the full UTI data set against a database of 18.8 million microbial sequences from UniProtKB (Release 2017_07), ProteoStorm completed in 1.79 CPU-days – a 485x speedup compared to the estimated 123.8 CPU-weeks for MS-GF+. Given these encouraging speedups and noting the extensive time requirements of other tools, we used a subset of 17 individuals (0.9 million spectra) to perform a systematic comparison of identified peptides and runtimes against MS-GF+, Comet (Eng, et al., 2013), and MSFragger (Kong, et al., 2017).

At 1% peptide-level FDR, MS-GF+ identified 12,139 peptides in an estimated CPU-22 weeks, Comet identified 9,341 peptides in an estimated CPU-10.7 weeks, and MSFragger identified 11,530 peptides in an estimated CPU-2.4 weeks. In contrast, ProteoStorm identified 13,550 peptides in 9.7 CPU-hours, a speedup by 382x, 186x, and 41x, respectively (Fig. 1d). ProteoStorm identified 96% (11,657) of the MS-GF+ peptides, as also 95.9% (10,331) of the peptides identified by at least two of the three tools (Fig. S3b). On the other hand, Comet identified 64.3% and 78.1% of peptides respectively, and MSFragger identified 75.9% and 93.5% of peptides respectively.

All tools had comparable identification rates, with ProteoStorm slightly higher at 5.98% in contrast to the range of 5.02% to 5.42% for other tools. The low identification rates were likely due to missing sequences in the initial pan-microbial database. To test this, we also searched a small wastewater metaproteomics dataset (150,216 spectra), where a specialized database (2.47Gb) had been constructed using the Graph2Pep/Graph2Prot approach (Tang, et al., 2016) on metagenomics data. ProteoStorm identified 10,633 peptides, or 95.9% (4,790) of the MS-GF+ peptides, in 52.6 minutes, while the Graph2Pep/Graph2Prot approach identified 8,740 peptides, or 83.6% (4,175) of the MSGF+ peptides, and took over 3.8 CPU hours (Fig. S1b). Graph2Pep/Graph2Prot reduced the initial search space by 124x, while ProteoStorm by 184x (Fig. S1a). Compared to searching against the UniProtKB database, searching against the metagenome derived database improved identification rate from 7.2% to 15.5%.

The speedup provided by ProteoStorm is further amplified by searching the full UTI data set against a comprehensive database (RefUP++; STAR Methods) containing 95 million sequences. When increasing both the number of spectra from 0.9M to 7.5M and the database size from UniProtKB (7.8 Gb) to RefUP++ (34.1Gb), ProteoStorm speedup over MS-GF+ increased from an estimated 382x to 953x (Fig. S3c). The reduction of the initial search space ranged from 34x to 76x depending on the dataset and database searched (Fig. S3a). Breaking-down ProteoStorm runtime by module shows that the increase is mostly due to a prolonged computation of *p*-values as the number of spectra increases (Fig. 1e). While there is also an increase in runtime when solely increasing the initial database size, this is mainly due to the stage-two database partitioning.

The initial UTI study (Yu, et al., 2015) restricted the search database to a manually-curated list of 20 genera, identifying 15 as being expressed in 110 individuals. Using an unbiased

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

search of the full UTI dataset against the RefUP++ database (2,259 genera) and a genera-restriction approach (STAR Methods), we identified 64 genera, including the previous 15 (Table S1). Out of 73,092 peptides, 28.5% (20,833) mapped uniquely to a single genus. Moreover, in 98 out of the 110 individuals, we found that every individual had either the same or a superset of the genera previously reported. Overall, we observed a complex pattern of polymicrobial expressions (Fig. 2). Using clusters supported by multiscale bootstrap resampling (Table S2), we could observe at least four bi-clusters. The most common uropathogen in UTI is *Escherichia coli*, followed by other gram-negative microbes, including *Klebsiella*, *Citrobacter*, *Pseudomonas*, and *Enterobacter* (Ronald, 2002). Individuals infected by these dominant uropathogens are present in clusters C-I (genera: 89au; individuals: 99au), including eight of the twelve individuals who were also clinically diagnosed with UTI (Yu, et al., 2017). Six of the eight individuals were aged 61 to 84, and exhibited polymicrobial UTI (p-UTI) with either *Escherichia* or *Klebsiella* dominant over the other (Table S1). Indeed, p-UTI is common among the elderly, and *Escherichia coli* and *Klebsiella* frequently co-occur in p-UTI (Laudisio, et al., 2015).

Individuals infected by obligate or facultative anaerobic pathogens that are frequently undetected using typical urine culture diagnostic methods are present in cluster B-III (genera: 97au; individuals: 94au) (Imirzalioglu, et al., 2008). Of note, individual #116 was clinically diagnosed with UTI, and while a previous study reported *Proteus mirabilis* as the major UTI causing pathogen, ProteoStorm identified the rare UTI pathogen *Propionimicrobium* (1,129 out of 3,531 PSMs) (Ikeda, et al., 2017) as the most abundant genera followed by *Facklamia* (544 PSMs), *Actinotignum* (538 PSMs), *Proteus* (424 PSMs), and other gram-positive bacteria (Table S1).

As the female urine specimens were clean catch samples, we cannot distinguish between true urethral/bladder colonization and possible contamination by microbes that are part of the vaginal and vulvar microbiota. In cluster A-II (individuals: 91au), which is dominated by *Lactobacillus*, we observed seven of twelve individuals who previously expressed high abundances of human epithelia cell proteins (Yu, et al., 2017). The dominance of Lactobacilli together with evidence of shed squamous epithelial cells could suggest a lack of an underlying pathology. In contrast, individuals in cluster A-IV (genera: 96a; individuals: 89au) revealed higher abundances of microbes known to contribute to bacterial vaginosis (BV), notably *Gardnerella*, *Prevotella*, *Atopobium*, *Sneathia* (Ma, et al., 2012; Onderdonka, et al., 2016). This data is consistent with BV as an underlying pathology related to these specimens.

As a UTI dataset is considered less complex, we searched a human infant gut metaproteomics dataset (2.4M spectra) against the RefUP++ database using the genera-restriction approach (STAR Methods). ProteoStorm identified 19 genera (Table S3) from 31,606 peptides mapping uniquely to a single genus (54.9% of all identifications), and detected similar shifts in microbial abundance across day of life 21, 34, and 50 when compared to the previous study (Xiong, et al., 2017) (Fig. S2a). Searching against the matched metagenome database, ProteoStorm identified 61,364 peptides, or 94.5% (53,847) of the MS-GF+ peptides (Fig. S2b) at an identification rate of 24.2%.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Metaproteomic studies based on database search approaches depend on the effective and efficient analysis of data. In clinical cases, the accurate identification of pathogens affects treatment options, and failure to detect certain microbes due to non-inclusiveness in the search database may lead to suboptimal treatment or misuse of antibiotics. The use of reference protein repositories is advantageous, but as they grow with increasing sequencing data, so does the need for efficient computational tools that analyze complex, multi-species communities. ProteoStorm is an ultrafast and highly sensitive tool for the metaproteomics community, and is available through GitHub (https://github.com/miinslin/ProteoStorm).

STAR Methods

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Vineet Bafna (vbafna@cs.ucsd.edu).

METHOD DETAILS

Data sources

Pan-microbial databases: The UniProtKB database (7.8Gb) in this study consists of 18,755,274 protein sequences from reference proteomes of 6,521 bacterial species (UniProt Release 2017.07). A comprehensive database (RefUP++; 34.1Gb; 2,259 genera) was constructed using 94,916,719 bacterial and fungal protein sequences from RefSeq (Release 85; ftp://ftp.ncbi.nlm.nih.gov/refseq/release/), and UniProt Pan Proteomes (Release 2017.12) and Reference Proteomes (Release 2017.07) (ftp://ftp.uniprot.org/pub/databases/uniprot/). We provide the RefUP++ database as a set of 58 files, accessible via Github. The fasta header information has been processed to include the name of the protein entry followed by the genus name of the organism and the taxonomy ID (see Genera-restriction approach).

Urinary Tract Infection (UTI) MS/MS dataset: To evaluate the performance of ProteoStorm, we used a dataset previously published in two related studies (Yu, et al., 2015; Yu, et al., 2017). LC-MS/MS Raw files of urine pellet (UP) samples from 110 suspected urinary tract infection (UTI) cases and five healthy individuals were provided by researchers at The J. Craig Venter Institute (JCVI) or downloaded from the PRIDE Database (PXD004713). As described previously (Yu, et al., 2017), the samples were collected for diagnostic purposes, considered medical waste, and de-identified prior to transfer to JCVI. Samples were run in replicate on an Ultimate 3000-nano LC and Q Exactive mass spectrometer system coupled via a FLEX nano-electrospray ion source (Thermo Scientific). RAW files were subjected to peak-picking (Vendor) and converted to Mascot generic format (MGF) format using MSConvert (ProteoWizard v3.0) (Chambers, et al., 2012).

Given the focus of ProteoStorm on metaproteomic data analysis, we performed an initial search of the full UTI dataset against the Human reference proteome (UP000005640; Release 2017.07) using MS-GF+ (variable modifications: protein N-terminal acetylation, and methionine oxidation; static modifications: cysteine carboxyamidomethylation), and excluded the 3,913,142 spectra that were identified as high-confidence human-matching

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

spectra (1% PSM-level FDR per MS/MS experiment) from the spectra partitioning step, resulting in 7,552,992 spectra for the full UTI dataset.

Human Infant Gut MS/MS dataset: To evaluate the performance of ProteoStorm on a more complex dataset, we searched a subset of a human infant gut metaproteomics dataset (MassIVE ID: MSV000080565). Spectra representing stool samples collected on day of life 21, 34, and 50 from infant 23 (healthy preterm infant with mild lung disease) were chosen based on the higher complexity of the samples as determined in the previous study (Xiong, et al., 2017). High-confidence human-matching spectra (135,380) were excluded from the spectra partitioning step, resulting in 2,360,544 spectra.

Wastewater MS/MS dataset: To evaluate the performance of ProteoStorm when searching against a matched metagenome database, we searched a dataset (150,216 spectra) representing oleaginous mixed microbial communities sampled from the surface of a biological wastewater treatment plant on Oct 12, 2011 (SD6; PeptideAtlas ID: PASS00577) (Muller, et al., 2014).

ProteoStorm core module 1: Data partitioning—ProteoStorm performs a fast, fully-tryptic search in the first stage followed by a semi-tryptic search in the second stage that is limited to proteins identified in the first stage. Each stage consists of three core modules: i) database and spectra partitioning, ii) ion-mass index-based peptide filtering, and iii) PSM *p*-value computation.

In the first module, *in silico* digested peptides and spectra are sorted and binned into n matching partitions based on parent mass. The theoretical mass of a peptide, p, with cysteine carboxyamidomethylation as a static modification, m_p is defined as the sum of the monoisotopic masses of residues in p, a hydrogen atom, a hydroxyl group, and a carbamidomethyl group for each cysteine in p. Given a distribution of theoretical masses for N peptides, we define partition i by a pair of indic (b_i, e_i) where the number of peptides satisfying $bi \leq m_p < e_i$ is $,N/n$ We provide mass distribution files for both the UniProtKB and RefUP++ databases, accessible via Github.

For spectra, to allow for at most one isotopic error of δ (i.e., 1.003355), and a precursor ion mass tolerance of ε (e.g., 10ppm), partition i is defined slightly differently as

$$b_i^s = b_i\left(1 - \left(\frac{\varepsilon}{10^6}\right)\right)$$

$$e_i^s = e_i\left(1 + \left(\frac{\varepsilon}{10^6}\right)\right) + \delta$$

Spectra partitioning: Given the interval pairs $\left(b_i^s, e_i^s\right)$ for all i, we assign spectra to n bins based on parent mass. The mass of a spectrum, s is defined as $m_s = M\,z - p^+\,z$ where M is

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

the precursor ion mass (in units of m/z), z is the charge state of the precursor ion, and p^+ is the mass of a proton. We assign spectrum s to each bin i if

$$b_i^s \leq m_s \leq e_i^s$$

To bin spectra efficiently, we sort them by their mass and use a merge operation to assign them to the appropriate bins. Spectra with less than 10 peaks are not included in the spectral partitioning process. As the distribution of spectra based on parent mass does not necessarily follow that of theoretical peptides, a maximum spectral partition size is enforced, creating multiple files for spectra belonging to bin i.

Database partitioning: ProteoStorm takes as input a set of database files, D, and assigns an index j to each file, where $0 \leq j < |D|$ Given D, files are read in batches of 1Gb, and protein sequences are stored in memory as a single string, T, where the character "$" indicates the start of a protein (an additional "$" character is appended at the end of the string to denote the end of the string). Each string T is subjected to *in silico* tryptic digestion, where peptides are limited to the standard 20 amino acids, isoleucine-transformed (i.e., leucine replaced with isoleucine), and allowed at most one missed cleavage site. If a peptide at the protein N-terminus, cleavage of its n-terminal methionine is allowed.

Each peptide, p, is stored in a string representation of its corresponding database partition i, and written to file when a buffer size of 600MB is reached. To map peptides back to proteins after stage one, we store additional information along with the sequence for each peptide p, including the index, j, of the originating database file, the protein index h within database file &›, and the pre- and post-amino acids. For a target-decoy approach (Elias & Gygi, 2007), protein sequences are reversed and concatenated in a similar manner to form string F. Origin from a decoy protein is indicated by prefix 'd_' for protein index h. For a semi-tryptic *in silico* digest, either the n-terminus or c-terminus of a fully-tryptic peptide is anchored and amino acids are removed from the opposite terminus to form semi-tryptic peptides.

As the same peptide, p, may result from the *in silico* digestion of one or more strings T (or F), ProteoStorm combines multiple mapping information from duplicate peptide entries in each partition i, and writes the mass of an unique peptide p, the peptide sequence, and the combined mapping information of p to create the final database partition. Additionally, ProteoStorm assigns a category to peptide p in the following order of preference: fully-tryptic/target, fully-tryptic/decoy, semi-tryptic/target, or semi-tryptic/decoy. The reason for this is twofold. First, a given spectrum s is always assigned the peptide sequence from its best scoring (s,p) pair, and given that fully-tryptic peptides score higher than semi-tryptic peptides during p-value computation, fully-tryptic peptides are preferred over semi-tryptic peptides. Second, in the case that a (s,p) pair with a peptide from a decoy sequence scores equally as well as one with a peptide from a target sequence, the latter is always preferred.

As an example, the following peptides have a mass withindatabase partition 0 in a tab separated format. The peptide sequence "AGGGGGGG" can be mapped to two proteins, specifically, the first occurrence is in the 2,936th entry of database file d_{4076}, or

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

UP000061569_69.fasta, and the 1,632nd entry in database file d_{2526} or UP000030518_1300345.fasta. The pre- and post-amino acid "R-" indicates that the peptide is a fully-tryptic protein c-term peptide.

In the second stage of ProteoStorm, the refined protein database (see Refined database creation) is provided as two files, a target sequence database and a decoy sequence database. A semi-tryptic *in silico* digest is performed, and database partitions are created as described previously, with the exception that sequences from the target database are not reversed to form string *F*. Instead, sequences from the decoy database are concatenated to form string *F*.

ProteoStorm core module 2: Peptide filtering—The second module of ProteoStorm computes the shared peak counts (SPC) between each spectrum and candidate peptides within its parent mass tolerance using an efficient ion-mass indexed data structure. For each database partition, peptide ions (b-/y-ions) with a maximum fragment ion charge of 2+ are used to construct an ion-mass indexing data structure, I, where each ion-mass index, $i \in I$, references a mass-sorted list of peptides containing a b-/y-ion within the mass tolerance of i (Fig. 1c). For each spectrum, s, peaks with an intensity less than 1.0 Q, where Q is the mean of the intensities of peaks at the bottom 25%, are considered noise and removed. Prominent peaks are defined as the top seven intense peaks in a window of 75 Daltons. For each prominent peak in a spectrum s, its corresponding ion-mass index i is accessed to retrieve a mass-sorted list of peptides, P, containing a matching ion. For all candidate peptides $p \in P$ that are within the parent mass tolerance of the spectrum and at most one isotopic error away from the spectrum parent mass, the shared peak count (SPC) of the spectrum-peptide pair (s,p) is incremented by one

To increase the efficiency of database search (filtering as many peptides as possible) without sacrificing sensitivity (retaining true positives), for each spectrum s, any candidate peptide with a SPC less than max($M_{\min}$, $M_{\max}$ −1) is filtered where

$$M_{max} = \max_{p \in P}(\mathrm{SPC}(s, p)).$$

We chose $M_{\min} = 7$ for fully-tryptic peptodes and $M_{\min} = 7$ for semi-tryptic peptides which works well for both the HCD fragmentation-based UTI dataset, the CID fragmentation-based human infant gut dataset, and the HCD fragmentation-based wastewater dataset. Determining $M_{\min}$ based on the length and mass of a peptide, the instrument type, and fragmentation method is optimal but requires further investigation.Retained spectrum-peptide pairs are subjected to further analysis in core module 3.

ProteoStorm core module 3: *p*-value computation—To accurately estimate the *p*-value of spectrum-peptide pairs and allow for computation of a false discovery rate, we created a modified version of MS-GF+ (v2018.04.09) (Kim & Pevzner, 2014) that directly calls its raw score computation and generating function approach. To identify high-confidence PSMs using the target-decoy approach, the peptide that achieves the highest MS-GF+ raw score for each spectrum s among the retained (s,p) pairs from core module 2 is selected for *p*-value computation using the generating function approach. Briefly, given an

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

SPC score of k for a spectrum-peptide pair (s,p), the generating function computes the probability of a randomly generated peptide (all residues independently and identically distributed) obtaining an SPC score of k or higher against spectrum s.

Refined database creation—Define a peptide group as an unmodified peptide sequence and the sum of its modifications (e.g., 'PEPTIDES+57.021464'). The p-value of a peptide group is equal to that of its best scoring (s, p) pair. For any p-value threshold, , letnumber of decoy groups whose p-value is ≤ θ. Similarly, let $t(\theta)$ denote the number of denote the number of target groups where p-value ≤ θ. A peptide-level false discovery rate (FDR) is thus defined as

$$\mathrm{FDR}(\theta) = \frac{f(\theta)}{t(\theta)}.$$

To determine high-confidence peptides, L, in the first stage of ProteoStorm, a script identifies peptide groups across all MS/MS experiments, and determines the p-value threshold for a peptide-level FDR of 5%. Peptide groups that pass the threshold are retained as high-confidence peptides $p \in L$. Subsequently, all proteins containing a peptide $p \in L$ are included in a refined protein database, D'. Using the protein mapping information retained during database partitioning, a script iterates through target peptides in L, writing both the protein sequence and its reverse sequence to D'. Redundant protein sequences, including those that only differ by protein name, are not allowed. A similar procedure is applied to high-confidence decoy peptides, with the exception that only decoy protein sequences are written to D'. Proposed by Bern and Kil (Bern & Kil, 2011), this method of generating a decoy database for a second stage search includes a conservative bias by including more decoy sequences than target. The more sophisticated version of this method, where the reversed sequences of target proteins are added to the decoy protein sequences until the total number of decoy sequences is equal to the number of target sequences was validated by Jeong et al. (Jeong, et al., 2012). The refined protein database is provided in the S1_OutputFiles directory as a combined target decoy database "RefinedProteinDB.fasta."

ProteoStorm Output—After completion of stage two, the identified PSMs (no FDR applied) are reported in the S2_OutputFiles directory as file "ProteoStorm_output.txt." For all analyses in this manuscript, we applied a 1% peptide-level FDR (pooled across all MS/MS experiments) for peptide identifications, and a 1% PSM-level FDR (per MS/MS experiment) for PSM identifications, where PSMs representing high-confidence peptides are reported.

ProteoStorm is designed to be flexible with regards to user needs. As an example, we used peptides from stage two of ProteoStorm to analyze metaproteomic samples at the genera level. Details on using ProteoStorm for other taxonomies and post-translational modifications are accessible via Github (Alternative configurations for ProteoStorm).

Genera-restriction approach—To identify microbial communities and infection patterns in the UTI dataset, we implemented a genera-restriction approach for ProteoStorm, where the pan-microbial database, D, includes only proteomes with available taxonomic

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

information at the genus level, and the refined protein database, *D'*, is based on high-confidence peptides identified in stage-one that uniquely map to a single genus among the restricted genera (see Refined database creation below). The three core modules of ProteoStorm are as described previously.

Database preprocessing: To determine the genus to which a proteome or protein entry belongs, a script parses NCBI taxonomy .xml files downloaded using the NCBI E-utilities API, and rewrites the fasta header information to include the name of the protein entry followed by the genus name of the organism and the taxonomy ID. For UniProt proteomes, file names already include a proteome identifier (UPID) that can be queried using the UniProt API to obtain its corresponding taxonomy ID. For RefSeq entries, taxonomy IDs are found in the RefSeq-release#.catalog file. Given that the UTI dataset represents human microbiome samples, we limited proteins to sequences belonging to genera from bacteria and fungi.

Refined database creation: To construct a genera-restricted protein database, high511 confidence peptides (1% peptide-level FDR), *C*, identified in stage-one of ProteoStorm are used to infer a set of genera, G_c, with at least one peptide $p \in H$ uniquely mapping to each genus *g*. Specifically, each genus, *g*, has a relative abundance score

$$A_g = \frac{S_g}{\sum_i^G S_g},$$

where S_g is the number of peptides in *C* that map uniquely to proteins belonging to genus *g*, and *G* is the set of all available genera. Target peptides in *C* contribute to the score of target genera, while decoy peptides in *C* contribute to the score of decoy genera. Genera with a higher A_g score than the most abundant decoy genera are included in the set of high-confidence genera, G_c. As abundance thresholds are largely dependent on the complexity of a microbial sample, we assumed that genera with a higher abundance than the most-abundant decoy genera are more likely to be present in the samples.

Given the set of high-confidence genera, G_c, and the list of high-confidence peptides (1% peptide-level FDR), *C*, a refined protein database is constructed in a similar manner to the procedure described previously. The additional criterion is that all proteins mapping to a peptide $p \in C$ must also belong to a genus $g \in G_c$.

To construct a refined protein database based on a different level of taxonomy, users can use the taxonomy ID provided for each protein entry in the RefUP++ database to query the NCBI taxonomy database (NCBI E-utilities API or taxonomy .xml files accessible via Github). A list of high-confidence taxa would be based on peptides mapping uniquely to that taxa (see Github: Alternative configurations for ProteoStorm).

Comparisons against conventional tools—For timing and identification comparisons of ProteoStorm against conventional database search tools, we required all processes to use a single core and less than 8Gb of RAM on an Intel® Core™ i7–6700K CPU (4 cores, 8

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

threads) with an installation of Windows 10 Pro. The UniProtKB database (7.8Gb) and 12+5 UTI dataset were used in the comparison. Given the engineering differences of MS-GF+ (v. 20161026) (Kim & Pevzner, 2014), Comet (2016.01 rev. 0) (Eng, et al., 2013), and MSFragger (20170103_v2) (Kong, et al., 2017), different approaches were implemented to satisfy the criteria mentioned above. Additionally, given the extensive time requirements of these tools, we estimated runtimes and obtained results through parallel runs on a Linux server or the local machine. In all searches, cysteine carboxyamidomethylation was included as a static modification. A 1% peptide-level FDR, as described previously, was applied using SpecEvalue for MS-GF+, e-value for Comet, and probabilities from running PeptideProphet (parameters: --clevel 0 --combine --decoyprobs --expectscore --ppm --accmass --nonparam) for MSFragger.

MS-GF+: Spectra were arbitrarily split into six files of size 800Mb and D was arbitrarily split into k database files of size 60Mb (132 database files for the UniProtKB database; 573 database files for the RefUP++ database). Spectra file #3 contained the closest number of spectra to the average number of spectra across all six files (157,945 spectra), and was searched against six database files with the highest number of fully-tryptic peptides. Runtime was estimated as the average runtime of the six database searches multiplied by the total number of spectra files and database files. Search parameters: -t 10ppm -tda 1 -m 3 -inst 3 -e 1 -ntt 1 -minLength 8 -maxLength 40 -thread 1. PSMs with peptides having greater than one miscleavage were removed after computing a 1% peptide-level and 1% PSM-level FDR.

Comet: A spectrum_batch_size of 25,000 was specified in the parameter file, and a combined spectra file was searched against the same six database files used to estimate runtime for MS-GF+. Runtime was estimated as the average of the six database searches, multiplied by the total number of database files. Default search parameters (comet.params.high-high) were used with the following exceptions:

decoy_search = 1

num_threads = 1

peptide_mass_tolerance = 10.00

num_enzyme_termini = 1

allowed_missed_cleavage = 1

no variable mods,

activation_method = HCD

digest_mass_range = 488.0 5700.0

clip_nterm_methionine = 1

spectrum_batch_size = 25000 decoy_prefix = XXX_

MSFragger: Spectra were arbitrarily split into five files of size 1000Mb and D was arbitrarily split into 1,583 target-decoy concatenated database files of size 10Mb. Spectra file

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

#1 contained the closest number of spectra to the average number of spectra across all five files (191,219 spectra). Runtime was estimated as the average of the database searches, multiplied by the total number of spectra files and database files. Default search parameters were used with the following exceptions:

num_threads = 1

precursor_mass_tolerance = 10.00

precursor_mass_units = 1

precursor_true_tolerance = 10.00

isotope_error = 0/1

num_enzyme_termini = 1

no variable modifications

digest_min_length = 8

digest_max_length = 40

digest_mass_range = 488.0 5700.0

max_fragment_charge = 3

minimum_peaks = 10

min_matched_fragments = 4 #recommended for narrow window search

minimum_ratio = 0.00

Comparison against Graph2Pep/Graph2Prot approach—To provide a fair comparison of ProteoStorm against MS-GF+ and the Graph2Pep/Graph2Prot approach (Tang, et al., 2016), we reconstructed the initial database (MG_SD6.500.faa and MG_SD6.contig-pep.fixedKR.fasta) used for the SD6 metaproteomics dataset by following the README file provided in the Graph2Pro GitHub repository (https://github.com/COL-IU/Graph2Pro). The reanalysis was limited to the SD6 dataset as the assembled contig file (SRR1544596-SD6MG-s2-63mer-k31-d1.contig) and edge file from graph (SRR1544596-SD6MG-s2-63mer-k31-d1.updated.edge) were only available for SD6 (http://darwin.informatics.indiana.edu/Dbgraph/example/). The following parameters were used for all searches utilizing MS-GF+: -t 15ppm -tda 0 -m 3 -inst 1 -e 1 -ntt 1 -minLength 8 -maxLength 40, with the exception of step 9 (second stage search) for the Graph2Pep/Graph2Prot approach, where -tda 1 was used. ProteoStorm parameters: --PrecursorMassTolerance 15 --FragmentMassTolerance 0.015 --InstrumentID 1 --FragmentMethodID 3.

For the Graph2Pep/Graph2Prot approach, we followed all steps until the eighth step, where an error was encountered in the Graph2Pro software. As such, to obtain final results for the Graph2Pep/Graph2Prot approach, we ran MS-GF+ on the provided second stage database (SD6_hybrid_DBGraphPep2Pro_5.fasta) using the parameters above. For the default MS-GF+ run, the SD6.mgf file was searched against a combined target-decoy database

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

(MG_SD6.contig-pep.fixedKR.fasta, MG_SD6.500.faa, and reversed protein sequences from MG_SD6.500.faa; 2.47 GB in size; 43,308,680 entries). Following the previous study (Tang, et al., 2016), we applied a 1% PSM-level FDR to obtain the final identifications.

Comparison against human infant gut dataset: To evaluate ProteoStorm on a more complex dataset, we searched the infant gut dataset (infant 23, day of life 21, 34, 50) against a matched metagenome database and compared to MS-GF+ results. As we already searched the dataset against a Human reference proteome (see Data sources) with contaminant sequences to remove high-confidence human-matching spectra, we removed Human and contaminant sequences from the metagenome database provided by the previous study (Xiong, et al., 2017) (Infant23_Human_Ref2011_IgA_contams.fasta), resulting in a microbial sequence database of 84,562 entries (27.5Mb). Parameters for MS-GF+ : -t 10ppm -m 1 -inst 1 -e 1 -ntt 1. Parameters for ProteoStorm: --PrecursorMassTolerance 10 --FragmentMassTolerance 0.6 --InstrumentID 1 --FragmentMethodID 1. A 1% peptide-level FDR was applied as described previously.

To identify the taxonomic composition of infant 23's gut microbiome, we searched the dataset against the RefUP++ database using the genera-restriction approach in ProteoStorm. A normalized genera matrix was constructed as previously described, with the following exception: Normalizing was based on a set of human proteins, , that were expressed in all samples collected from infant 23 (i.e., day 21 run 1, day 21 run 2, day 34 run 1, day 34 run 2, day 50 run 1, day 50 run2).

QUANTIFICATION AND STATISTICAL ANALYSIS

Genera matrix—After the second stage of ProteoStorm, a genera matrix is constructed, where each row is a genus, *g*, each column is an individual, and values are normalized relative abundances in spectral counts, averaged across technical replicates. While there isn't an optimal way to normalize across experiments, we normalized based on spectral counts for a set of human proteins, *H*, that were expressed in all healthy cohort samples (representative of a baseline expression level for each MS/MS experiment). To identify genera, we performed exact matching of high-confidence peptides (1% peptide-level FDR), *C'*, to protein sequences in the pan-microbial database, *D*, using the PeptideMatchCMD_1.0.jar tool (Chen, et al., 2013). The normalized relative abundance for a genus *g* in a MS/MS experiment, *E*, is defined as

$$\left(G_a\right)_N = \frac{G_a}{\alpha_E},$$

where G_a is the number of PSMs passing a 1% PSM-level FDR that belong to a peptide $p \in H'$ that maps uniquely to genus *g*. The normalization factor, α_E, is the sum of PSMs passing a 1% PSM-level and 1% peptide-level FDR that map to a protein in , divided by the number of proteins in *H* expressed in experiment *E*. If a peptide matches to more than one protein, the number of PSMs is divided by the number of matched proteins.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Hierarchical clustering—As suggested by literature (Clarke & Green, 1988), the pseudo count-adjusted ($+1\times10^{-30}$) genera matrix was square root transformed prior to computing Bray-Curtis dissimilarity matrices (vegdist function in R package vegan v2.4.6) for both the genera (rows) and individuals (columns). Unsupervised hierarchical clustering of the dissimilarity matrices was performed separately (R function hclust, method: ward.D2). For better visualization, the square root transformed genera matrix was further $\log_{10}$ transformed prior to heatmap construction (heatmap.2 function in R package gplots v3.0.1).

Clusters were evaluated using an edited version of the R package pvclust (v2.0.0) (Table S2). The source code was downloaded from https://github.com/cran/pvclust, and line 382 in pvclust-internal.R was changed from "dist(t(x),method)" to "vegdist(t(x), method = "bray", binary=FALSE)", where vegdist is the function from R package vegan (v2.4.6) that computes a dissimilarity index (Bray-Curtis) and returns a distance object. The bootstrap sample size, nboot, was set at 100,000. Clusters with a probability value (approximately unbiased (AU) *p*-value) greater than 0.9 were considered strongly supported by multiscale bootstrap resampling.

DATA AND SOFTWARE AVAILABILITY

ProteoStorm is a collection of scripts written in python (v2.7), that calls an executable compiled from C++ scripts for peptide filtering (core module 2) as well as a java .jar executable from a modified version of MS-GF+ for raw score and *p*-value computations (core module 3). ProteoStorm is available at https://github.com/miinslin/ProteoStorm.

The LC-MS/MS urinary pellet dataset is available on MassIVE (MSV000082031) in RAW file format and peak-picked MGF file format.

TABLE FOR AUTHOR TO COMPLETE

Please upload the completed table as a separate document. **Please do not add subheadings to the Key Resources Table**. If you wish to make an entry that does not fall into one of the subheadings below, please contact your handling editor. (**NOTE:** For authors publishing in Current Biology, please note that references within the KRT should be in numbered style, rather than Harvard.)

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors would like to thank Seungjin Na, Ben Pullman, Mingxun Wang, and Seong Won Cha for helpful discussions, Seungjin Na for providing c++ code for computing theoretical ion m/z values, and Christopher Wilkins for helpful discussions on the usage of the MS-GF+ codebase. D.B., M.S.L., and V.B. were supported by grants from the NIH (P-41-RR24851 and 1R01GM114362).

References

Bern M, and Kil YJ (2011). Comment on "Unbiased Statistical Analysis for Multi-Stage Proteomic Search Strategies". J Proteome Res 10, 2123–2127. [PubMed: 21288048]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Burke MC, Mirokhin YA, Tchekhovskoi DV, Markey SP, Heidbrink Thompson J, Larkin C, and Stein SE (2017). The Hybrid Search: A Mass Spectral Library Search Method for Discovery of Modifications in Proteomics. J Proteome Res 16, 1924–1935. [PubMed: 28367633]

Chambers MC, Maclean B, Burke R, Amodei D, Ruderman DL, Neumann S, Gatto L, Fischer B, Pratt B, Egertson J, et al. (2012). A cross-platform toolkit for mass spectrometry and proteomics. Nat Biotechnol 30, 918–920. [PubMed: 23051804]

Chatterjee S, Stupp GS, Park SK, Ducom JC, Yates JR, 3rd, Su AI, and Wolan DW (2016). A comprehensive and scalable database search system for metaproteomics. BMC Genomics 17, 642. [PubMed: 27528457]

Chen C, Li Z, Huang H, Suzek BE, Wu CH, and UniProt C (2013). A fast Peptide Match service for UniProt Knowledgebase. Bioinformatics 29, 2808–2809. [PubMed: 23958731]

Clarke KR, and Green RH (1988). Statistical design and analysis for a 'biological effects' study. Marine Ecology Progress Series 46, 213–226.

Elias JE, and Gygi SP (2007). Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat Methods 4, 207–214. [PubMed: 17327847]

Eng JK, Jahan TA, and Hoopmann MR (2013). Comet: an open-source MS/MS sequence database search tool. Proteomics 13, 22–24. [PubMed: 23148064]

Ikeda M, Kobayashi T, Suzuki T, Wakabayashi Y, Ohama Y, Maekawa S, Takahashi S, Homma Y, Tatsuno K, Sato T, et al. (2017). Propionimicrobium lymphophilum and Actinotignum schaalii bacteraemia: a case report. New Microbes New Infect 18, 18–21. [PubMed: 28491325]

Imirzalioglu C, Hain T, Chakraborty T, and Domann E (2008). Hidden pathogens uncovered: metagenomic analysis of urinary tract infections. Andrologia 40, 66–71. [PubMed: 18336452]

Jagtap P, Goslinga J, Kooren JA, McGowan T, Wroblewski MS, Seymour SL, and Griffin TJ (2013). A two-step database search method improves sensitivity in peptide sequence matches for metaproteomics and proteogenomics studies. Proteomics 13, 1352–1357. [PubMed: 23412978]

Jeong K, Kim S, and Bandeira N (2012). False discovery rates in spectral identification. BMC Bioinformatics. 13 Suppl 16, S2.

Kim S, and Pevzner PA (2014). MS-GF+ makes progress towards a universal database search tool for proteomics. Nat Commun 5, 5277. [PubMed: 25358478]

Kong AT, Leprevost FV, Avtonomov DM, Mellacheruvu D, and Nesvizhskii AI (2017). MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. Nat Methods 14, 513–520. [PubMed: 28394336]

Laudisio A, Marinosci F, Fontana D, Gemma A, Zizzo A, Coppola A, Rodano L, and Antonelli Incalzi R (2015). The burden of comorbidity is associated with symptomatic polymicrobial urinary tract infection among institutionalized elderly. Aging Clin Exp Res 27, 805–812. [PubMed: 25916348]

Ma B, Forney LJ, and Ravel J (2012). Vaginal microbiome: rethinking health and disease. Annu Rev Microbiol 66, 371–389. [PubMed: 22746335]

Muller EEL, Pinel N, Laczny CC,Hoopmann MR, Narayanasamy S, Lebrun LA, Roume H, Lin J, May P, Hicks ND, Heintz-Buschart A, Wampach L, Liu CM; Price LB, Gillece JD, Guignard C, Schupp JM, Vlassis N, Baliga NS, Moritz RL, Keim PS, and Wilmes P (2014) Community-integrated omics links dominance of a microbial generalist to fine-tuned resource usage. Nat Commun 5, 5603. [PubMed: 25424998]

Onderdonk AB, Delaney ML, and Fichorova RN (2016). The Human Microbiome during Bacterial Vaginosis. Clin Microbiol Rev 29, 223–238. [PubMed: 26864580]

Rinner O, Seebacher J, Walzthoeni T, Mueller LN, Beck M, Schmidt A, Mueller M, and Aebersold R (2008). Identification of cross-linked peptides from large sequence databases. Nat Methods 5, 315–318. [PubMed: 18327264]

Ronald A (2002). The etiology of urinary tract infection: traditional and emerging pathogens. Am J Med 113 Suppl 1A, 14S–19S. [PubMed: 12113867]

Stein SE (1995). Chemical substructure identification by mass spectral library searching. J Am Soc Mass Spectrom 6, 644–655. [PubMed: 24214391]

Tang H, Li S, and Ye Y (2016). A Graph-Centric Approach for Metagenome-Guided Peptide and Protein Identification in Metaproteomics. PLoS Comput Biol 12, e1005224. [PubMed: 27918579]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Xiong W, Brown CT, Morowitz MJ, Banfield JF, and Hettich RL (2017). Genome-resolved metaproteomic characterization of preterm infant gut microbiota development reveals species-specific metabolic shifts and variabilities during early life. Microbiome. 5, 72. [PubMed: 28693612]

Yu Y, Sikorski P, Bowman-Gholston C, Cacciabeve N, Nelson KE, and Pieper R (2015). Diagnosing inflammation and infection in the urinary system via proteomics. J Transl Med 13, 111. [PubMed: 25889401]

Yu Y, Sikorski P, Smith M, Bowman-Gholston C, Cacciabeve N, Nelson KE, and Pieper R (2017). Comprehensive Metaproteomic Analyses of Urine in the Presence and Absence of Neutrophil-Associated Inflammation in the Urinary Tract. Theranostics 7, 238–252. [PubMed: 28042331]

Zhang X, Ning Z, Mayne J, Moore JI, Li J, Butcher J, Deeke SA, Chen R, Chiang CK, Wen M, et al. (2016). MetaPro-IQ: a universal metaproteomic approach to studying human and mouse gut microbiota. Microbiome 4, 31. [PubMed: 27343061]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Highlights:

- Open-source software for ultrafast large-scale metaproteomics database search
- Two to three orders-of-magnitude speedup over popular tools, at high sensitivity
- Reveals rare pathogens in reanalysis of UTI dataset.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

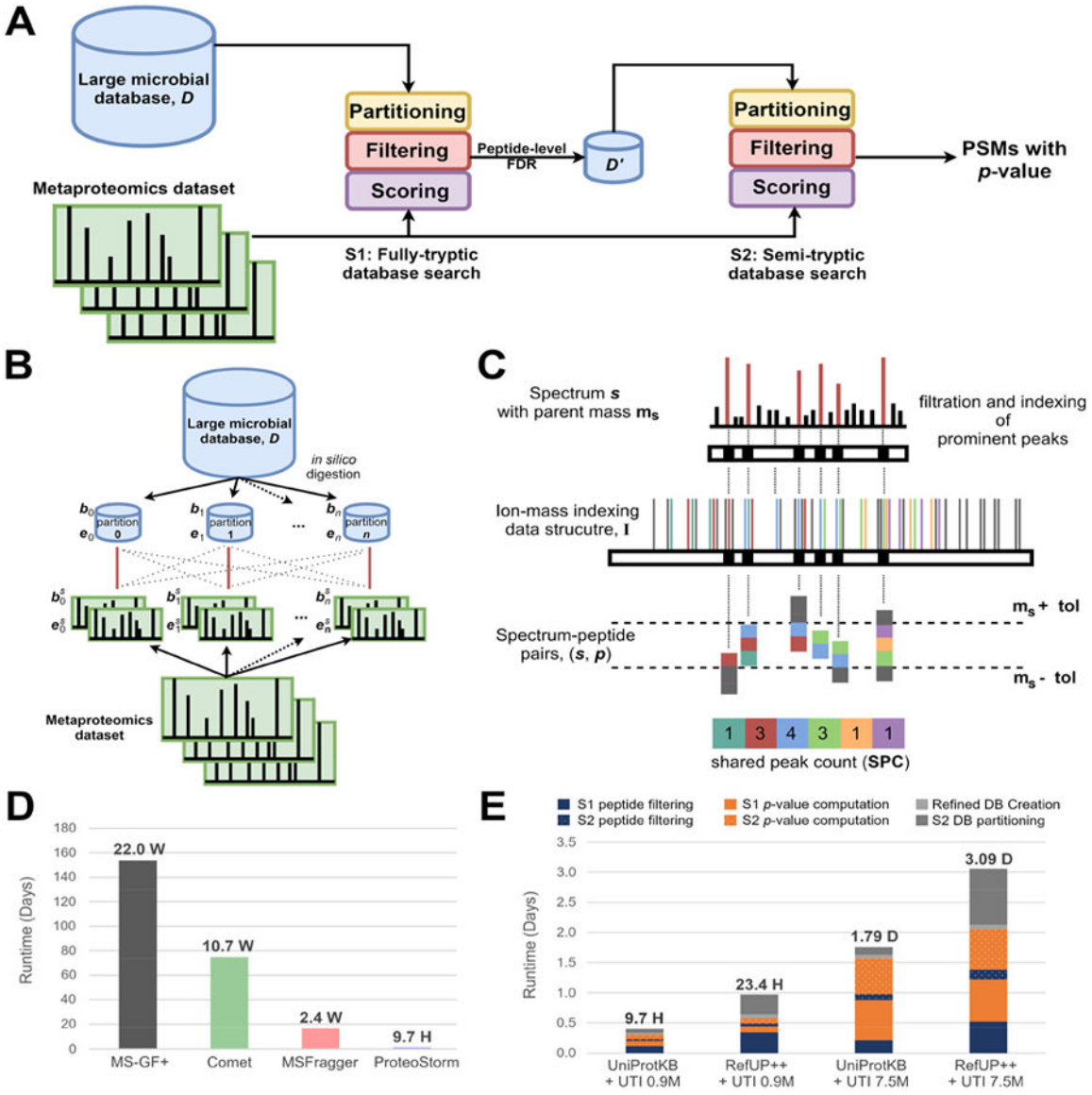


Figure 1. ProteoStorm search framework: performance and scalability

(a) Each stage of ProteoStorm is composed of three modules: i) data partitioning, ii) peptide filtering, and iii) *p*-value computation. Identifications from a fully-tryptic search are used to construct a refined protein database for a semi-tryptic search. PSMs with *p*-values are reported. **(b)** In the first module, database and spectra partitioning dramatically reduces search space from all peptides (black dotted lines) to peptides within spectra parent mass ranges (red vertical lines). **(c)** In the second module, spectrum-peptide pairs are filtered based on shared counts of prominent spectral peaks (red) to b-/y-ions of peptides (numbers within blocks) using an ion-mass indexing data structure. Each colored block represents a unique peptide within the parent mass tolerance of a given spectrum. **(d)** 946,845 spectra were searched against the UniProtKB database using ProteoStorm, MSGF+, Comet, and MSFragger. ProteoStorm required 9.7 CPU-hours, while other tools required CPU-weeks to

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

complete. **(e)** Breakdown of ProteoStorm runtime by module. S1 and S2 represent the two different stages of ProteoStorm. See also Figure S3.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

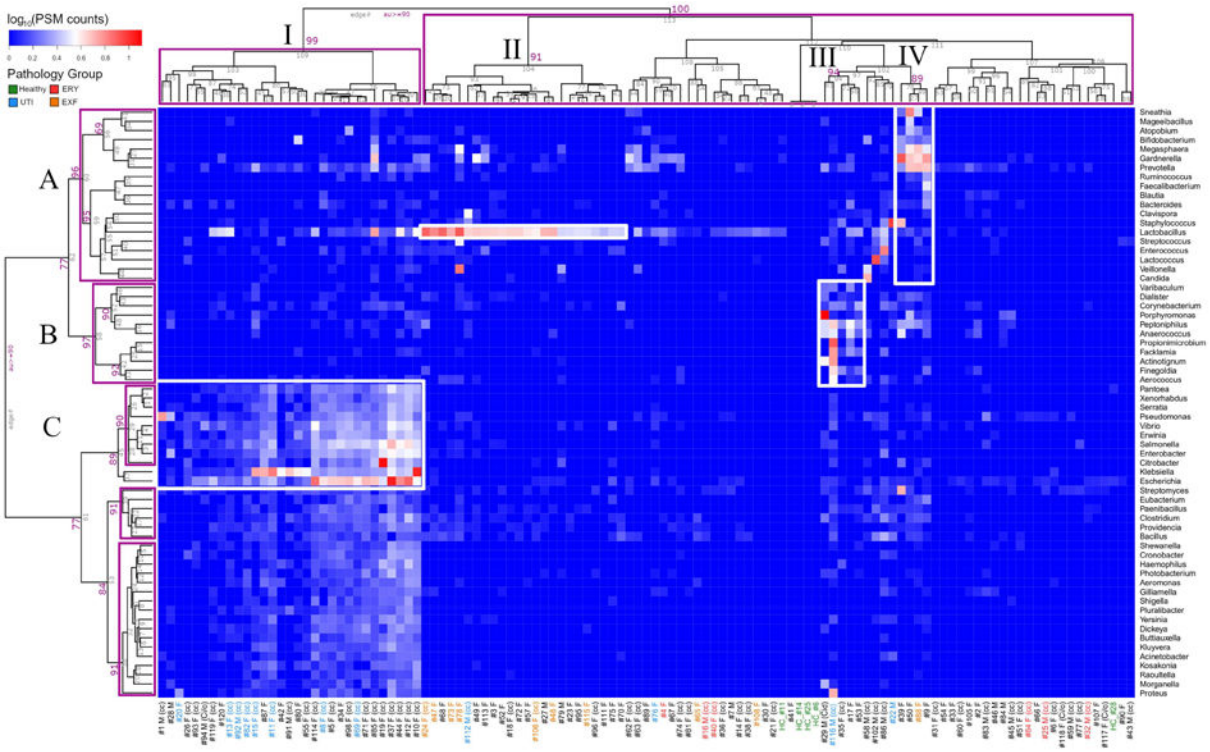


Figure 2. ProteoStorm identifies bi-clusters of individuals with similar microbial compositions

Searching the full UTI dataset against the RefUP++ database (2,259 genera) using a genera-restriction approach, ProteoStorm identified 64 genera. Out of 73,092 peptides, 28.5% (20,833) mapped uniquely to a single genus. Four bi-clusters (white boxes) were inferred from clusters with an approximately unbiased (au) *p*-value greater than 0.90 (magenta boxes), indicating a complex pattern of polymicrobial expression, including sub-types of urinary tract infections, cases of bacterial vaginosis, and evidence of no underlying disease. Pathology groups: Healthy, ERY (erythrocyte/vascular injury), EXF (exfoliation of squamous epithelial and urothelial cells), and UTI (urinary tract infection). See also Table S1 and Table S2.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Mass (Da)	Peptide Sequence	Peptide Index	Mapping Information
488.1979	R.AGGGGGGG.-	0	4076|2936|R-;2526|1632|R-
502.2136	R.GAGGAGGG.-	XXX_1	4640|d_2272|R-
502.2136	R.GGGGGGAA.-	2	6530|897|R-

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
Urinary Tract Infection (UTI) MS/MS dataset	Yu, et al.,2015; Yu, et al.,2017	MasslVE ID: MSV000082031
Human Infant Gut MS/MS dataset	Xiong, et al., 2017	MasslVE ID: MSV000080565
SD6 Wastewater MS/MS dataset	Muller, et al.,2014	PeptideAtlas ID: PASS00577
UniProt Pan proteomes	ftp://ftp.uniprot.org/pub/databases/uniprot	Release 2017.12
UniProt Reference proteomes	ftp://ftp.uniprot.org/pub/databases/uniprot	Release 2017.07
RefSeq protein sequences	ftp://ftp.ncbi.nlm.nih.gov/refseq/release	Release 85
Software and Algorithms		
ProteoStorm	This paper	https://github.com/miinslin/ProteoStorm
MSConvert	Chambers, et al.,2012	v3.0
MS-GF+	Kim & Pevzner, 2014	v20161026
Comet	Eng, et al., 2013	2016.01 rev. 0
MSFragger	Kong, et al., 2017	20170103_v2
Graph2Pep/Graph2Pro	Tang, et al., 2016	https://github.com/COL-IU/Graph2Pro
pvclust	https://github.com/cran/pvclust	v2.0.0
vegan	https://CRAN.R-project.org/package=vegan	v2.4.6
dendextend	https://CRAN.R-project.org/package=dendextend	v1.7.0
gplots	https://cran.r-project.org/package=gplots	v3.0.1