

Tennessee State University

Digital Scholarship @ Tennessee State University

Chemistry Faculty Research

Department of Chemistry

2-12-2013

A Novel Algorithm for Validating Peptide Identification from a Shotgun Proteomics Search Engine

Ling Jian
Vanderbilt University

Xinnan Niu
Vanderbilt University


Zhonghang Xia
Western Kentucky University

Parimal Samir
Vanderbilt University

Chiranthani Sumanasekera
Vanderbilt University

See next page for additional authors

Follow this and additional works at: <https://digitalscholarship.tnstate.edu/chemistry-faculty>

 Part of the [Biochemistry Commons](#), and the [Organic Chemistry Commons](#)

Recommended Citation

L. Jian, X. Niu, Z. Xia, P. Samir, C. Sumanasekera, Z. Mu, J.L. Jennings, K.L. Hoek, T. Allos, L.M. Howard, K.M. Edwards, P.A. Weil, and A.J. Link "A Novel Algorithm for Validating Peptide Identification from a Shotgun Proteomics Search Engine" *Journal of Proteome Research* 2013 12 (3), 1108-1119 DOI: 10.1021/pr300631t

This Article is brought to you for free and open access by the Department of Chemistry at Digital Scholarship @ Tennessee State University. It has been accepted for inclusion in Chemistry Faculty Research by an authorized administrator of Digital Scholarship @ Tennessee State University. For more information, please contact XGE@Tnstate.edu.

Authors

Ling Jian, Xinnan Niu, Zhonghang Xia, Parimal Samir, Chiranthani Sumanasekera, Zheng Mu, Jennifer L. Jennings, Kristen L. Hoek, Tara Allos, Leigh M. Howard, Kathryn M. Edwards, P. Anthony Weil, and Andrew J. Link



Published in final edited form as:

J Proteome Res. 2013 March 1; 12(3): 1108–1119. doi:10.1021/pr300631t.

A Novel Algorithm for Validating Peptide Identification from a Shotgun Proteomics Search Engine

Ling Jian^{1,@}, Xinnan Niu[@], Zhonghang Xia³, Parimal Samir⁴, Chiranthani Sumanasekera^{*}, Mu Zheng, Jennifer L. Jennings, Kristen L. Hoek, Tara Allos, Leigh M. Howard[#], Kathryn M. Edwards[#], P. Anthony Weil[†], and Andrew J. Link²

¹Department of Pathology, Microbiology and Immunology, Vanderbilt University School of Medicine, Nashville, TN, School of Mathematical Sciences, Dalian University of Technology, Dalian, China

³Department of Mathematics and Computer Science, Western Kentucky University, Bowling Green, KY

⁴Department of Biochemistry, Vanderbilt University School of Medicine, Nashville, TN

^{*}Department of Molecular Physiology and Biophysics, Vanderbilt University School of Medicine, Nashville, TN

[#]Department of Medicine, Vanderbilt University School of Medicine, Nashville, TN

Abstract

Liquid chromatography coupled with tandem mass spectrometry has revolutionized the proteomics analysis of complexes, cells, and tissues. In a typical proteomic analysis, the tandem mass spectra from a LC/MS/MS experiment are assigned to a peptide by a search engine that compares the experimental MS/MS peptide data to theoretical peptide sequences in a protein database. The peptide spectra matches are then used to infer a list of identified proteins in the original sample. However, the search engines often fail to distinguish between correct and incorrect peptides assignments. In this study, we designed and implemented a novel algorithm called De-Noise to reduce the number of incorrect peptide matches and maximize the number of correct peptides at a fixed false discovery rate using a minimal number of scoring outputs from the SEQUEST search engine. The novel algorithm uses a three step process: data cleaning, data refining through a SVM-based decision function, and a final data refining step based on proteolytic peptide patterns. Using proteomics data generated on different types of mass spectrometers, we optimized the De-Noise algorithm based on the resolution and mass accuracy of the mass spectrometer employed in the LC/MS/MS experiment. Our results demonstrate De-Noise improves peptide identification compared to other methods used to process the peptide sequence matches assigned by SEQUEST. Because De-Noise uses a limited number of scoring attributes, it can be easily implemented with other search engines.

Keywords

proteomics; mass spectrometry; bioinformatics; support vector machines; peptide spectrum match; database search engine; validation

²Corresponding Author: Andrew J. Link, Ph.D., Dept of Pathology, Microbiology and Immunology, Vanderbilt University School of Medicine, Nashville, TN 37232, andrew.link@vanderbilt.edu, Tel: 6153436823, Fax: 6153437392.

[@]These authors contributed equally to this work

Supporting Information Available: This material is available free of charge via the Internet at <http://pubs.acs.org>.

Introduction

Liquid chromatography coupled with tandem mass spectrometry (LC/MS/MS) offers the promise to comprehensively identify and quantify the proteome of complexes, cells and tissues. The large numbers of peptide spectra generated from LC/MS/MS experiments are routinely searched using a search engine against theoretical fragmentation spectra derived from target databases containing either protein or translated nucleic acid sequences. It is typically assumed that a peptide spectrum match (PSM) for each MS/MS spectrum is contained in the sequence database. In a typical peptide identification procedure, PSMs are ranked according to either a cross correlation, a statistical score, or a probability that the match between the experimental and theoretical is correct and unique. Only those PSMs with the highest scores or most significant probabilities are reported as correct. However, this approach often falsely identifies the peptides. In reality, more than 50% of PSMs initially assigned by database search engines, such as SEQUEST, MASCOT, and X! TANDEM are incorrect^{1,2}. As a result, the accuracy of database search results is often evaluated by searching a decoy protein database to identify the false discovery rate (FDR)²⁻⁸. Decoy databases contain either reversed or randomly shuffled protein sequences derived from the target protein database. The database search engine assigns an observed spectrum to either a target or a decoy sequence. The assignment of a peptide from a decoy database to an experimental spectrum is considered incorrect because it is assumed that there is no such peptide sequence in reality. The target-decoy database search also indicates the quality or reliability of the target PSMs. Nonetheless, the target PSMs are not all correct due to either the poor quality of the experimental MS/MS data, the absence of the sequence in the database, or unexpected amino acid modifications. As a consequence, a fraction of the target PSMs from the search engine is false positive. Hence, manual or computational approaches are essential to validate target PSMs after a search engine-protein database analysis of LC/MS/MS data.

SEQUEST is one of the most widely used approaches for automatically assigning the observed spectra generated from a LC/MS/MS experiment to peptide sequences in a sequence database⁹. However, the original SEQUEST algorithm does not include a statistical method to determine the specificity of peptide-spectrum matching. An early approach to identify correct target PSMs uses empirical score filters set at defined score thresholds to validate PSMs from a SEQUEST search^{10,11}. PSMs above the defined thresholds are accepted as correct, while those below are assumed to be incorrect. The empirical score filters are not always easily defined due to the multiple scoring metrics derived from SEQUEST scores and the variable quality of the mass spectrometry data. Also, the accuracy of the validated PSMs derived from an empirical scoring filter varies with the type of mass spectrometer used.

Different approaches have been developed to validate peptide assignments¹². One of the most commonly used computational tools is PeptideProphet, which uses a Bayesian statistical algorithm to convert SEQUEST scores into probabilities¹³. With PeptideProphet, conditional probabilities for the PSMs are computed by the expectation maximization (EM) method, using the assumption that the PSM data are drawn from a mixture in which the distribution of the correct and incorrect PSMs follows a prescribed Gaussian model. A list of PSMs above a predefined posterior probability is reported¹³. An updated version of PeptideProphet utilizing a semi-supervised technique was recently developed¹⁴. It integrates the EM algorithm with a decoy database search strategy to build a classifier based on a Bayesian probability model.

To provide a more efficient way of evaluating SEQUEST outputs, MacCoss and Noble's group employed support vector machine (SVM), a powerful classification technique^{16,17}, to

classify correct and incorrect PSMs after a database search^{15,16,17}. SVM-based classification is a supervised learning approach that uses training data to build a model and assign a label to each data point. They used this approach to create the algorithm called Percolator to directly distinguish correct from incorrect PSMs⁴. The goal of Percolator is to increase the number of correct target PSMs reported at a minimal FDR or q-value¹⁸. Starting with a small set of trusted correct PSMs and a set of incorrect PSMs from searching a decoy database, Percolator iteratively adjusts the learning model to fit the dataset by ranking high-confidence PSMs higher than decoy peptide matches. With the given q-value, this approach iteratively trains the classifier and eventually results in a classifier that has an improved ability to distinguish correct from incorrect PSMs.

In this study, we have developed a novel algorithm called De-Noise for statistical validation of correct target PSMs identified by SEQUEST. De-Noise uses a SEQUEST search of a concatenated database containing both target and decoy proteins. It uses the decoy PSMs as incorrect references for measuring the reliability of the correct target PSMs. The De-Noise algorithm is a continuous refining process by which the incorrect target PSMs or noisy PSMs are sequentially eliminated. First, it computes the distance of every target PSM to the centroid of the decoy PSMs. With the assumption that the target PSMs close to the centroid of the decoy PSMs are incorrect or noise, they are eliminated based on a defined ratio. The remaining dataset provides a set of PSMs with improved quality for building an SVM-based decision function to refine the target and decoy PSM. Using a given false positive rate (FPR), De-Noise distinguishes the correct from incorrect target PSMs using two rounds of SVM-based decision functions and refinement. Specifically, the lowest scoring target PSMs are discarded from the dataset based on the scores derived from SVM-based decision functions until the FPR is reached. Next, the algorithm sorts the remaining PSMs based on the expected protease digestion patterns into three categories: canonical, half-canonical, and non-canonical. It assigns the protease digestion categories an expectation factor based upon the expected distribution of the three categories. With the expectation factor and a score ($PSM_{evaluator}$) derived from normalized SEQUEST's Xcorr and DeltaCn scores, De-Noise further refines the PSMs to eliminate the incorrect target PSMs.

Our results demonstrate De-Noise has increased sensitivity and specificity for validating PSMs after a SEQUEST search compared to both PeptideProphet and Percolator. To evaluate the performance of De-Noise, we used LC/MS/MS datasets generated from various control and biological samples run on different mass spectrometers. The mass spectrometers had a wide range of mass accuracies, resolutions, and user-defined capabilities for selecting the precursor ions to fragment. The low and high data quality datasets were used to develop and evaluate our De-Noise algorithm. Our results demonstrate that De-Noise validates more correct target PSMs under a series of fixed FDRs compared to PeptideProphet and Percolator. The target PSMs validated by these two algorithms extensively overlaps De-Noise's validated PSMs. These results demonstrate De-Noise can increase the number of validated target PSMs.

Materials and Methods

Reagents

Universal Proteomics Standard Set (UPS1) was purchased from Sigma (St. Louis, MO). Partisphere strong cation exchange (SCX) material was purchased from Whatman International Ltd. Jupiter 3 μ C18 300A, reverse phase (RP) material was purchased from Phenomenex (Torrance, CA). Formic acid and HPLC-grade acetonitrile were obtained from Fisher Scientific (Pittsburgh, PA). Trypsin was purchased from Promega (Madison, WI). PEEK tubing, sleeves, microtee and microcross were obtained from Upchurch Scientific

(Oak Harbor, WA). Fused silica capillaries were purchased from Polymicro Technologies (Phoenix, AZ).

Sample Preparation and LC/MS/MS Analysis

UPS1—UPS1 was solubilized in water, reduced with dithiothreitol (DTT), alkylated with iodoacetamide (IAA), and trypsin digested as previously described¹⁹. The tryptic peptides were analyzed with RP microcapillary LC/nanoESI/MS/MS. Briefly, a fritless, microcapillary 100- μ m-inner diameter column was packed with 9 cm of Jupiter C18 RP material. A 0.5 pmol aliquot of the trypsin-digested UPS1 was loaded onto the RP column equilibrated in buffer A (0.1% formic acid, 5% acetonitrile). The column was placed in line with an LTQ linear ion trap mass spectrometer (ThermoFisher). The sample was eluted using a 60-min linear gradient from 0 to 60% buffer B (0.1% formic acid, 80% acetonitrile) at a flow rate of 0.3 μ l/min. During the gradient, the eluted ions were analyzed by one full precursor MS scan (400–2000 m/z) followed by MS/MS scans on the five most intense precursor ions detected in the precursor MS scan while operating under dynamic exclusion.

Gcn4—Affinity preparation of the *S. cerevisiae* Gcn4 complex and its MudPIT analysis using an LCQ quadrupole ion trap mass spectrometer (ThermoFisher) have been previously described¹⁹.

Tal08—*S. cerevisiae* transcription complexes were prepared from yeast YTT3675 cells using the Tal08 minichromosome²⁰ and Dynal beads (Invitrogen) cross-linked to anti-Flag M2 antibody (Sumanasekera *et al.* manuscript in preparation). For Tal08 complexes, 20 ng of BSA was added to the Tal08 sample as an internal standard. The Tal08 sample was reduced with DTT, alkylated with IAA, and trypsin digested as previously described¹⁹. The desalted tryptic peptides were analyzed using a 11-step MudPIT experiment on a LTQ-Orbitrap XL (ThermoFisher). A 100 μ M ID fused silica microcapillary packed with 3 cm of Partisphere SCX material was coupled to a 12 cm long pulled fused silica capillary column packed with Jupiter C18 RP material. For the Mudpit run, the salt steps were 25 mM, 50 mM, 75mM, 100 mM, 150mM, 200 mM, 250 mM, 300 mM, 500 mM, 750 mM and 1M ammonium acetate. A 100 min linear RP gradient from 5% to 45% Buffer B was used for each salt step. Buffer A was 0.1% formic acid in HPLC-grade water and Buffer B was 0.1% formic acid in acetonitrile. A precursor ion scan was performed in the Orbitrap with preview mode and monoisotopic precursor selection (MiPS) enabled. The top 10 precursors ions based on intensity were fragmented in the ion trap using 35% normalized collision energy. Dynamic exclusion was enabled for 180 s with a repeat count of 1 for a 30s duration, a list size of 500, and an exclusion mass tolerance of 10 ppm.

PBMC—Human peripheral blood mononuclear cells (PBMC) were obtained from fresh venous blood using a Ficoll gradient protocol (Hoek *et al.*, manuscript in preparation). PBMCs were lysed in 50% trifluoroethanol (TFE) in 50 mM triethylammonium bicarbonate (TEAB) essentially as described²¹. The total protein content was quantified using a Bradford assay. The PBMCs lysate was reduced with DTT and alkylated with IAA. The sample was diluted 1:5 with 50mM TEAB to make the final concentration of TFE <10% and digested with trypsin as previously described¹⁹. The desalted sample was analyzed using a LTQ-Orbitrap XL and a LTQ-Orbitrap Velos (ThermoFisher). In the LTQ-Orbitrap XL analysis, 6-step MudPIT experiments were performed using either MiPS or MiPS-off. A 100 μ M ID fused silica microcapillary packed with 3 cm of Partisphere SCX material was coupled to a 12 cm long 100 μ m ID fritless pulled fused silica microcapillary capillary column packed with Jupiter C18 RP material. For both the MiPS and MiPS-off experiments, the MudPIT salt steps were 50 mM, 100 mM, 200 mM, 300 mM, 500 mM and 1M ammonium acetate. A 100 min RP linear gradient from 5% to 45% Buffer B was used for

each step. Buffer A was 0.1% formic acid in HPLC-grade water and Buffer B was 0.1% formic acid in acetonitrile. A precursor ion scan was performed in the Orbitrap with preview mode enabled. The top 10 precursors ions based on intensity were fragmented in the ion trap using 35% normalized collision energy. Dynamic exclusion was enabled for 180 s with a repeat count of 1 for a 30 s duration and a list size of 500. The exclusion mass tolerance was set to 10 ppm for MiPS. For MiPS-off, the exclusion mass width was set between 1.5 and 2.5. For the LTQ-Orbitrap Velos analysis, 11-Step MudPIT experiments were performed with either MiPS or MiPS-off, similar to Orbitrap XL experiments. The MudPIT salt pulses used were 25 mM, 50 mM, 75mM, 100 mM, 150mM, 200 mM, 250 mM, 300 mM, 500 mM, 750 mM and 1M ammonium acetate. A 100 μ M ID fused silica microcapillary packed with 3 cm of PartiSphere SCX material was coupled with a fritless, microcapillary 100- μ m-inner diameter column packed with 20 cm of Jupiter C18 RP material. A 90 min linear gradient from 2% to 40% Buffer B was used for each salt step. Buffer A was 0.1% formic acid in HPLC-grade water and Buffer B was 0.1% formic acid in acetonitrile. A precursor ion scan was performed in the Orbitrap with preview mode enabled. The top 16 precursor ions based on intensity were fragmented in the dual pressure ion trap with 35% normalized collision energy. For both MiPS and MiPS-off, dynamic exclusion was enabled for 15 s with a repeat count of 1 set for a 10 s duration, a list size of 500, and an exclusion mass width set between 0.5 and 0.5.

PMN: Polymorphonuclear cells—(PMN) were obtained from fresh venous human blood using a Ficoll gradient protocol (Hoek *et al.*, manuscript in preparation). PMNs were lysed in 50% TFE in 50 mM HEPES buffer as essentially described²¹. The total protein content was quantified using a BCA assay. The PMN lysate was reduced with DTT and alkylated with IAA. The sample was diluted 1:5 with 50 mM HEPES to make the final concentration of TFE <10% and digested with trypsin as previously described¹⁹. A 6-step MudPIT experiments was performed on a LTQ mass spectrometer (ThermoFisher) as previously described¹⁹.

Tandem MS Data analysis with SEQUEST

The RAW files generated from the LCQ, LTQ and Orbitrap LC/MS/MS experiments were converted to dat or mzXML formats with the program ReadW. The MS/MS spectra were extracted from the mzXML file using the program MzXML2Search and the data were analyzed using the SEQUEST algorithm to search either a Sigma48, *S. cerevisiae* (SGD_2010) or human Uniprot (uni280910) target and decoy concatenated protein database^{22,23}. All decoy databases were created by reversing the sequences in the target databases. For Percolator, separate target and decoy searches were performed. For all data processing, a static modification of 57.021464 for cysteine was used. All SEQUEST searches were performed with no enzyme specificity.

Analysis of SEQUEST Database Search Result Using PeptideProphet

To validate the PSMs identified by SEQUEST, the SEQUEST outputs from the LC/MS/MS experiments were loaded into the Trans Proteomic Pipeline V.4.0.2 (TPP). The search outputs were converted to pep.XML format files and analyzed by the TPP program PeptideProphet²⁴. Validation of the PSMs was performed by testing a range of probability filters until the desired FDR was reached. The pep.XML output file from PeptideProphet was converted to a CSV format. The CSV file was parsed with the in-house Perl script digest4peptide.pl, to sort the validated PSMs into lists of full-, half-, or non-canonical tryptic peptide consensus sequences.

Analysis of SEQUEST Database Search Using ATP

The SEQUEST *.out files were concatenated by an in-house Perl script `grab_files_threaded.pl` to generate a merged *.outs file. The concatenated *.outs file was parsed and loaded into an Oracle relational database using the in-house Perl script `concurrent_loading.pl` and processed and analyzed using BIGCAT/ATP^{25,26}. Two previously described filters, with low and high thresholds, were used to validate PSMs^{10,27}. The low-threshold filter for PSMs was set with cutoff values of $X_{corr} = 1.5$ for +1 charge state spectra, $X_{corr} = 2$ for +2 spectra, and $X_{corr} = 2$ for +3 spectra. Only fully-canonical PSMs were accepted¹⁰. For a high-threshold filter²⁷, PSMs with a +1 charge state were valid if they were fully-canonical and had an $X_{corr} > 1.9$. PSMs with a +2 charge state were valid if they were fully-canonical or half-canonical and had X_{corr} ranges between 2.2 and 3.0. PSMs with a +2 charge state and an $X_{corr} > 3.0$ were valid regardless of the PSM's protease consensus pattern. Finally, +3 peptides were valid if they were fully- or half-canonical and had an $X_{corr} > 3.75$. The filtered outputs from both filters were stored in CSV-formatted files and analyzed using Microsoft Excel.

Analysis of SEQUEST Database Search Using Percolator

The target and decoy SEQUEST outputs from the LC/MS/MS experiments were converted to a merged file in SQT format²⁸ using an in-house modified version of the program `Unitemare.pl` (<http://fields.scripps.edu/downloads.php>). The UNIX `sed` utility was used to remove the header information of the converted SQT files. Two entries, H SQT Generator SEQUEST and H SQTGeneratorVersion2.7, were added as headers to the SQT files so that they can be analyzed by Percolator. The SEQUEST target and decoy search results in SQT format were loaded into Percolator. A range of q-values were tested until the desired FDR was reached. The outputs were stored in tab delimited format. The outputs were parsed by the in-house Perl script `get_digest4percolator.pl` to sort the validated PSMs based into list of full-, half-, or non-canonical tryptic peptide consensus sequences.

Analysis of SEQUEST Database Search using the De-Noise Algorithm

The SEQUEST result in *.out format were converted to Microsoft Excel format and processed by the De-Noise algorithm implemented with Matlab version 7.8.0.347 running on a Dell T410 with the Windows Server 2008R2 Standard operating system, 32 GB of RAM, and an Intel® Xeon® CPU at 2.27GHz. Two support libraries and packages LibSVM library²⁹ and SKMsmo³⁰ needed for De-Noise decision function calculation were installed on the Dell T410³⁰. The Matlab functions `tic` and `toc` were used to measure the De-Noise execution time running on the T410 machine. The validated peptides generated by the De-Noise algorithm were exported in Microsoft Excel format.

Results and discussion

Generation of Test Datasets

One goal in developing the De-Noise algorithm was to create a PSM validation tool unbiased in terms of sample, type of mass spectrometer, or mass spectrometry method. Therefore we used seven LC/MS/MS datasets generated from a variety of control and experimental protein samples analyzed on different mass spectrometers. First, we used a prepared mixture of 48 known human proteins (UPS1), which allowed us to unambiguously identify incorrect and correct PSMs. Second, we used affinity purified yeast Gcn4 and Tal08 minichromosome complexes and Ficoll-purified human PBMCs, which were authentic biological samples and contained both expected and unexpected proteins. The Tal08 sample contained bovine serum albumin as an internal standard. All samples were trypsin-digested prior to LC/MS/MS analysis.

The Gcn4 and UPS1 samples were analyzed on LCQ and LTQ mass spectrometers, respectively. To maximize the number of precursor ions fragmented, these ion trap datasets measured the precursor ion masses at low resolution and mass accuracy³¹. The Tal08 and PBMC samples were analyzed on LTQ-Orbitrap instruments. The Orbitrap's high resolution and mass accuracy measurement of the precursor ions' m/z ratios allowed us to use monoisotopic precursor selection (MiPS) to select only peptide-like precursors for fragmentation^{34,32,33}. For the seven experiments, we searched the LC/MS/MS data against concatenated target and decoy databases using the SEQUEST algorithm with no protease enzyme specificity⁹. The top scoring SEQUEST PSM for each MS/MS spectrum was used for all downstream data validation. For spectra matched to peptide sequences in the target database, we assumed they were true hits and treated them as correct PSMs. Spectra matched to peptide sequences in the decoy database were considered false hits and were treated as incorrect PSMs. Based on this assumption, the FDR was computed using the equation below^{35,36}

$$FDR=2 \times D_n/(D_n+T_n)$$

where D_n is the number of the spectra matched to decoy peptide sequences and T_n is the total number of the PSMs matched to target peptide sequence. Table 1 summarizes the different samples, types of mass spectrometers, precursor ion selection methods, the number of MS/MS spectra collected, and the unfiltered SEQUEST search results using concatenated target and decoy protein databases. These datasets and PSMs were used for the design and evaluation of the De-Noise algorithm.

Variation in the Datasets

We observed several effects of the type of MS/MS analysis on the datasets. First, we compared the datasets obtained from the LCQ and LTQ to those obtained from the LTQ-Orbitrap XL and LTQ-Orbitrap Velos. There were distinct differences in the FDRs and decoy PSMs/total PSMs ratios when we compared the UPS1 and Gcn4 results to the Tal08 and PBMC results (Table 1). The LCQ and LTQ data had higher calculated FDRs and decoy PSM/total PSMs ratios compared the Orbitrap data. Compared to the LCQ and LTQ instruments, the Orbitrap mass spectrometers have higher mass accuracy and resolution and different precursor methods for selecting and analyzing precursor ions. Second, we found that the precursor selection feature (MiPS) of the Orbitrap instruments influenced the results. With MiPS, there were significantly fewer PSMs, target PSMs, and decoy PSMs compared to with MiPS-off (Table 1). Although the precursor selection method significantly influenced the total number of PSMs, there was little variation in the decoy PSMs/total PSMs ratios and the calculated FDRs if MiPS or MiPS-off was used. We postulated that these variations in the quality of the acquired mass spectrometry data needed to be taken into consideration in the design of the De-Noise algorithm.

Creating an Improved Dataset for the Decision Function

In a typical binary classification of correct and incorrect PSMs, the target PSMs are labeled as correct or +1, and decoy PSMs are labeled as incorrect or -1. The classifier learns from the training dataset to assign either +1 or -1 class labels to PSMs. However, a large number of PSMs assigned as target PSMs by SEQUEST are actually incorrect². These mislabeled target PSMs should be assigned to the incorrect class. If the mislabeled target PSMs were discarded from the PSMs dataset, we would create a better target PSM dataset to generate an improved decision function. A challenge was how to eliminate the incorrect target PSMs from the correct target PSMs.

The initial step in De-Noise is to cleanse the dataset of incorrect or noisy target PSMs. The PSM data were represented as vectors based on the attributes from five SEQUEST scores: *Xcorr*, *DeltaCn*, *SPrank*, *Ions*, and *Calc-neutral-pep-mass*. To avoid attributes with larger values dominating ones with smaller values, we normalized each of the original SEQUEST scores by using the following equation

$$x_nor = x_raw - (\text{mean of } x_raw) / (\text{std } x_raw)$$

where x_nor is the normalized SEQUEST score, x_raw is the original SEQUEST score, std x_raw is the standard deviation of the original SEQUEST score. Next, each target PSMs is classified as incorrect or correct by computing its distance to the centroid of the decoy PSMs which belongs to the -1 class. Specifically, given a set S with m PSMs, S consisted of $m^+ + 1$ and $m^- - 1$ PSMs. Let $S^+ = [x^+_1, x^+_2, \dots, x^+_{m^+}]$ be the set of +1 PSMs and $S^- = [x^-_1, x^-_2, \dots, x^-_{m^-}]$ be the set of -1 PSMs. We first apply multiple kernel methods to map the PSM data points using the function φ into feature space where the target and decoy can be separated more easily³⁷. The centroid of the -1 PSMs in the feature space denoted by x^-_C is computed using the equation below

$$x^-_C = \frac{1}{m^-} \sum_{i=1}^{m^-} \varphi(x_i)$$

where φ is the mapping function mapping PSM data points to feature space in which the inner product $\langle \varphi(x), \varphi(y) \rangle$ can be computed through the kernel function³⁷. Next, we computed the distance between each target PSM data point x_j and x^-_C in the feature space as follows

$$\begin{aligned} d(x_j, x^-_C) &= \left\| \varphi(x_j) - \frac{1}{m^-} \sum_{i=1}^{m^-} \varphi(x_i) \right\| = \left(\varphi(x_j) - \frac{1}{m^-} \sum_{i=1}^{m^-} \varphi(x_i), \varphi(x_j) - \frac{1}{m^-} \sum_{i=1}^{m^-} \varphi(x_i) \right)^{\frac{1}{2}} \\ &= \left(k(x_j, x_j) - \frac{2}{m^-} \sum_{i=1}^{m^-} k(x_j, x_i) + \frac{1}{m_-^2} \sum_{i=1}^{m^-} \sum_{k=1}^{m^-} k(x_i, x_k) \right)^{\frac{1}{2}} \end{aligned}$$

where $k(x_i, x_j)$ is the kernel function.

All +1 target PSMs with distances to the centroid of the -1 decoy PSMs less than a specific threshold are assumed to be incorrect and are discarded from the target PSM dataset. Let $d(x^+_i, x^-_C)$ be the distance from x^+_i to x^-_C in the feature space. For a threshold d_0 , x^+_i were selected as incorrect and should be removed if $d(x^+_i, x^-_C) < d_0$. In practice, it was a challenge to choose the optimal threshold d_0 because correct target PSMs could be discarded if d_0 is set too stringently. If it is too permissive, a large number of incorrect PSMs would remain in the dataset.

Instead of using a distance threshold, we set the number of discarded incorrect target PSMs by applying the ratio θ using the equation below.

$$\theta = \text{assumed false positive PSMs} / \text{total PSMs}$$

Previous studies have shown 10-50% of target PSMs identified by a database search engine are correct². The dilemma was how to determine θ to generate a target PSMs dataset cleansed of incorrect PSMs without throwing out correct target PSMs.

To guide our determination of θ , we took advantage of the distinct differences in the decoy PSMs/total PSMs ratio between the unfiltered LCQ and LTQ datasets and those derived from the Orbitrap datasets (Table 1). We assumed that the precursor ion spectra collected using the Orbitrap are of higher quality compared to the datasets obtained using LCQ and LTQ ion traps. A majority of MS/MS spectra (>60%) collected from the Orbitrap instruments were assigned to target PSMs, whereas only ~50% of the spectra collected from the LCQ and LTQ were assigned to target PSMs. We therefore inferred that if the decoy PSMs/total PSMs ratio is 40%, θ needs to be set to a larger value to cleanse more incorrect target PSMs. However, if the decoy PSMs/total PSMs ratio was <40%, we needed to set θ at a smaller value. Based on previous studies, θ was set at 0.1 for the UPS1 and Gcn4 datasets². This resulted in the elimination of additional incorrect target PSMs. For the Tal08 and PBMC datasets, we empirically optimized the values for θ by testing a θ range of 0.01 to 0.1. We found a θ of 0.03 resulted in the optimal number of incorrect target PSMs being cleansed from the target PSMs while keeping the distribution of fully-, half-, and non-canonical PSMs consistent with the expected ratios (Fig S1).

The De-Noise algorithm was designed to automatically determine the value of θ based on the decoy PSMs/total PSMs ratio from the unfiltered SEQUEST results. With the selected θ , the incorrect target PSMs were iteratively eliminated based on the distance of the target PSMs to the centroid of decoy PSMs. Target PSMs with the shortest distance to the centroid of the decoy PSMs are discarded first. This elimination of target PSMs considered noisy continued until the eliminated PSMs/total PSMs ratio satisfied θ .

Refining the PSM Datasets Using SVM-based Decision Functions

Finding an efficient decision function to calculate the decision score for each PSM was a critical step in the De-Noise algorithm. A kernel-based method provides a powerful learning tool for datasets with nonlinear structures and is adaptable to a variety of data types. The kernel method works by mapping data points within the vector space into a feature space where they can be easily separated. With the kernel method, we could combine different mappings by the sum of corresponding kernel matrices to provide complementary views of the data. In order to precisely characterize the relationships of each pair of PSMs, we tested the Polynomial, Gaussian, and Laplace kernel functions. Because it gave the greatest separation between target and decoy PSMs (Fig. S2), we selected Gaussian kernels computed with different weights using the equation

$$K = \sum_{i=1}^m \mu_i K^i$$

where K is the combination of individual Gaussian kernels, K^i , $i=1, \dots, m$, and μ_i are the corresponding weights. In our experimental studies, the kernel width of the individual

Gaussian kernel was chosen as 1, 0.5, and 0.2 respectively, and the weights μ_i were learned by using the SKMsmo software package³⁰.

After the noisy target PSMs are discarded from the original target PSMs dataset based on θ , two rounds of data refining with SVM-based decision functions are performed to separate the correct from the incorrect PSMs. In the first round, the updated target PSMs dataset S^+ and the decoy PSMs dataset S^- are combined into set S_0 to build the first SVM decision function. The target PSMs $x^+_i \in S^+$ are treated as incorrect PSMs if their decision function $f(x^+_i) \leq 0$ where $f(x^+_i)$ was determined by the SVM learning model. These incorrect target PSMs are discarded from S_0 to generate S_1 containing the remaining PSMs.

We observed that a subset of S_1 's target PSMs had $f(x^{Z^+}_i)$ scores $<$ decoy PSMs $f(x^-_i)$ scores. We assumed some of these PSMs to be incorrect and targeted them in a second SVM decision function and refinement. The refined target PSM dataset S_1 is used to build a second decision function. To remove the incorrect target PSMs, we applied a new parameter γ using the equation below.

$$\gamma = \text{number of retained decoy PSMs in } S_1 / \text{Total number of decoy PSMs in } S_1$$

In this second round, $f(x^-_{\gamma^-})$ is the score of the lowest scoring decoy PSM that was retained. By comparing $f(x^+_i)$ to $f(x^-_{\gamma^-})$, the target PSM x^+_i is discarded if $f(x^+_i) \leq f(x^-_{\gamma^-})$. De-noise iteratively tests a range of γ until the desired FDR is reached. As a result, a second set of incorrect target PSMs are cleansed to generate S_2 . In the refining processes, a default slack penalty parameter of 1 is used.

Refining Target PSMs using Proteolytic Peptide Patterns

After the refining steps, De-Noise used proteolytic patterns to validate the PSMs from S_2 . We categorized PSMs in the seven datasets into three groups based on their protease digestion patterns; fully-canonical, half-canonical, and non-canonical. Table 2 shows the majority of the PSMs representing fully-canonical peptides were assigned to target peptides. Only a very small number of fully-canonical PSMs were assigned to decoy peptides. These results indicated that the fully-canonical target PSMs are more likely to be correct matches and should be retained. However, the decoy/target ratio showed the number of half- and non-canonical target and decoy PSMs assigned to the LCQ and LTQ datasets were almost equal (Table 2). This strongly implied that a higher percentage of false positives were present in half- and non-canonical target PSMs compared to the fully-canonical PSMs. From this observation, we reasoned that a higher proportion of incorrect target PSMs from these two categories needed to be ultimately eliminated. We observed the Orbitrap datasets had similar decoy/target ratios for fully-canonical and non-canonical PSMs. However, the Orbitrap datasets had a lower decoy/target ratio for the half-canonical target PSMs compared to the half-canonical decoy PSMs from the LCQ and LTQ datasets. We reasoned that a higher percentage of the half-canonical target PSMs from the Orbitrap datasets should be retained compared to data from the LCQ and LTQ.

In a proteolytic digest using a site-specific protease, a majority of peptides are canonical followed by half canonical and non-canonical peptides. Therefore, a greater weight is applied to canonical peptides compared to half-canonical and non-canonical peptides. Using De-Noise, all fully-canonical peptides generated after the second refining were retained in the LCQ, LTQ and Orbitrap datasets. However, the distribution of half- and non-canonical PSMs coming from De-Noise were significantly different compared to the distributions generated by PeptideProphet and Percolator. To correct for this result, the half- and non-

canonical PSMs from the 2nd round of SVM-based refining were filtered depending on the type of mass spectrometer used. We developed an approach that aims to remove half- and non-canonical PSMs after the 2nd refining while optimizing both the number of validated PSMs and the distribution of half- and non-canonical PSMs. First, for the PSMs in dataset S_2 , we calculated a $PSM_{evaluator}$ value using the equation

$$PSM_{evaluator} = Xcorr_{Nor} + DeltaCn_{Nor}$$

where $Xcorr_{Nor}$ is the normalized Xcorr and $DeltaCn_{Nor}$ is the normalized DeltaCn described earlier. We used these two attributes because they have been previously shown to contribute the most towards measuring the accuracy and uniqueness of a PSM, respectively^{9,11,15}. Next we

τ_{half} = the number of half-canonical PSMs accepted / Total number of half-canonical PSMs in S_2

τ_{non} = the number of non-canonical PSMs accepted / Total number of non-canonical PSMs in S_2

Using a range of τ_{half} and τ_{non} , we developed an iterative approach to select the τ values that optimized the balance between the total number of validated PSMs and the distribution of half- and non-canonical PSMs. We tested a range of τ_{half} and τ_{non} to achieve an equivalent distribution of half- and non-canonical validated PSMs as reported by PeptideProphet and Percolator (Table S1 and S2). The values for τ_{half} and τ_{non} for the LCQ/LTQ datasets were determined to be 0.12 and 0.005, respectively (Table S1). For Orbitrap datasets, τ_{half} and τ_{non} were determined to be 0.5 and 0.005 (Table S2), respectively. With the τ_{half} and τ_{non} values, the half and non-canonical PSMs were discarded from the 2nd round of refining based on the $PSM_{evaluator}$ score. Finally, the pseudocode for the entire De-noise algorithm is summarized in Fig. 1.

Evaluating De-Noise's Performance

To evaluate the performance of De-Noise, we compared its performance against the PeptideProphet and Percolator algorithms for validating SEQUEST target PSMs^{10,11}. First, we measured De-Noise's runtime using the UPS1 and PBMC Orbitrap Velos MiPS-off datasets. It took ~43 s for De-Noise to process the UPS1 data, which was the smallest dataset and ~4082 s to validate the PBMC Orbitrap Velos MiPS-off data which was the largest. We found the runtimes were very similar to comparable approaches¹³⁻¹⁵. Second, we compared the number of PSMs identified by three algorithms. Table 3 shows De-Noise validated more PSMs than the semi-supervised PeptideProphet and Percolator learning approaches and two Xcorr filtering approaches (Low- & High- Stringency). Since the approaches using learning algorithms to validate PSMs from SEQUEST had the highest performance (Table 3), we focused our evaluation of De-Noise compared to PeptideProphet and Percolator.

To compare the validated PSMs from the De-Noise, PeptideProphet, and Percolator, we looked at the overlapping PSMs. Fig. 2 and S3 shows that the majority of De-Noise validated PSMs were also validated by PeptideProphet and Percolator. Table 4 shows a numerical summary of the overlapping validated PSMs from the three approaches. For example, for UPS1 dataset, 94% of the PSMs validated by PeptideProphet overlapped with De-Noise while 90% of the PSMs validated by Percolator were validated by De-Noise.

Similar patterns were seen in the other datasets in which De-Noise shared more validated PSMs with PeptideProphet than with the Percolator (Table 4, Fig. 2, and Fig. S3).

To show that our approach to remove half- and non-canonical target PSMs after the refining generated a similar distribution compared to PeptideProphet and Percolator, we compared the categorized overlapping outputs from the three approaches for the UPS1, Gcn4, and Tal08 datasets (Table 5 and 6). The validated PSMs and the overlapped PSMs from all three approaches for UPS1 and Gcn4 showed a similar distribution pattern. The number of validated fully-canonical PSMs was the largest class followed by the half-canonical and non-canonical PSMs, respectively (Table 5 and 6). The considerable overlap of validated PSMs from De-Noise, PeptideProphet, and Percolator (Table 5) and the similar distributions of PSMs (Table 6) showed De-Noise's approach to retain the most significant half- and non-canonical PSMs was valid.

We compared De-Noise, PeptideProphet, and Percolator using different FDR values. Fig. 3 shows the number of validated PSMs from the seven datasets using a series of FDRs. The performance of a validation approach is better if it validates more target PSMs compared to another approach with the same FDR. The plots in Fig. 3A and 3B demonstrated the De-Noise validated more target PSMs compared to PeptideProphet and Percolator. From the Gcn4 dataset, De-Noise validated approximately 2.8% and 13.2 % more PSMs than PeptideProphet and Percolator, respectively. Likewise, from the UPS1 dataset, De-Noise identified about 31% and 73% more PSMs than PeptideProphet and Percolator, respectively. For the Tal08 dataset, the validated PSMs increased 30% and 29% using De-Noise compared to PeptideProphet and Percolator, respectively. A similar pattern is seen in PBMC datasets (Fig. 3D-G). We observed that De-Noise consistently outperforms Peptide Prophet and Percolator in terms of the number of target PSMs validated at a given FDR.

Evaluating De-Noise for Sensitivity and Specificity

From an applied mathematics point of view, distinguishing correct from incorrect PSMs can be treated as a two-class classification problem, in which a classifier labels the PSMs as either true or false^{4,38}. There are four possible outcomes from a binary classifier. If the classifier validates a PSM as true and the spectrum is also matched to a target peptide sequence, then it is called a true positive (TP). However, if the classifier validates a PSM as true but the spectrum was matched to a decoy peptide sequence, then it is said to be a false positive (FP). Conversely, if the classifier assigns a PSM as false and the spectrum was matched to a decoy peptide sequence, then a true negative (TN) has occurred. Finally, if the classifier assigns a PSM as false and the spectrum was matched to a target sequence, a false negative (FN) has occurred. The true positive rate (TPR) is defined as the ratio of the target PSMs validated by a classifier to the total number of target PSMs.

$$TPR=TP/(TP+FN)$$

The false positive rate (FPR) is the ratio of the number of decoy PSMs falsely identified as a target PSMs to the total number of decoy PSMs.

$$FPR=FP/(FP+TN)$$

The performance of a binary classification method is measured using two statistical parameters: sensitivity and specificity. Sensitivity is equal to the TPR and reflects the classifier's capability to correctly validate target PSMs from a pool of target PSMs.

Specificity is equal to 1- FPR and measures the frequency at which the classifier correctly validates decoy PSMs from the total pool of decoy PSMs. The overall performance of the classifier can be represented with a receiver operating characteristic (ROC) curve³⁹, which plots the true positive rate (sensitivity) versus the false positive rate (1-specificity). Each point on the ROC curve represents sensitivity/specificity. When two classifiers are compared, the classifier with the higher sensitivity at a given specificity is considered the better classifier.

The performance of the De-Noise, PeptideProphet, and Percolator approaches was evaluated by using ROC plots³⁹ (Fig. 4). In general, the classifier with ROC plot closest to the left-hand border is considered the most robust. The robustness of a classifier declines as its curve gets closer to the 45 degree diagonal of the ROC space. Another index to assess the robustness of a classifier from a ROC plot is the area under the curve (AUC). The larger the area covered, the more robust is the classifier. For example, we calculated the AUC for the UPS1 dataset using the trapezoid rule for an FPR range from 0.0008 to 0.0074⁴⁰. For this FPR range, the AUC calculated for the three approaches shows De-Noise is the more robust approach (De-Noise (0.0006), PeptideProphet (0.00038), and Percolator (0.00025 (Fig. 4B). For the other six datasets, the AUC for De-Noise was consistently larger compared to PeptideProphet and Percolator showing that De-Noise was more robust (Fig. 4A-G). Using the ROC curves, we compared the sensitivity for the three approaches. For the seven datasets, De-Noise consistently had the higher TPR compared to PeptideProphet and Percolator in the FPR range 0.01 to 0.05 (Fig 4A-G).

Evaluating De-Noise's Performance using an Independent Dataset

Finally, to test De-Noise's performance in validating PSMs on a dataset not used in its optimization, the De-Noise algorithm was applied to a human PMNs extract run on a LTQ ion trap mass spectrometer. The results from the SEQUEST search for the PMN LTQ data set are shown in Tables 1 and 2. Table 3 shows De-Noise validated more target PSMs compared to other validation approaches.

In summary, we have developed a highly sensitive and specific algorithm to validate PSMs from the SEQUEST search engine. The novel De-Noise algorithm first uses a data cleaning step based on the distance of the target PSMs to the centroid of the decoy PSMs to remove noisy, incorrect PSMs from the target PSMs. Second, De-Noise performs two rounds of data refining using SVM-based decision functions to validate correct target PSMs. Finally, the algorithm uses proteolytic information and the quality of the mass spectrometry data to perform a final validation. Using a variety of datasets based on different samples, mass spectrometers, and popular validation approaches, we show the De-Noise algorithm has improved sensitivity and specificity in the 1-5% FDR range that is commonly used to report the accepted peptide sequences from tandem mass spectrometry search engines.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by NIH grant GM064779 and Vanderbilt University School of Medicine IDEAS Program grant 1-04-066-9530. This project has been funded in part with Federal funds from the National Institutes of Allergy and Infectious Disease, National Institutes of Health, Department of Health and Human Services, under Contract No. 272200800007C and the Vanderbilt Clinical and Translational Science Award grant NIH RR024975. The software in this manuscript is available upon request. We thank Drs Tony Weil and Kristen Hoek for allowing us to cite unpublished results.

References

1. Elias JE, Haas W, Faherty BK, Gygi SP. Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations. *Nat Methods*. 2005; 2:667–675. [PubMed: 16118637]
2. Elias JE, Gygi SP. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods*. 2007; 4:207–214. [PubMed: 17327847]
3. Peng J, Elias JE, Thoreen CC, Licklider LJ, Gygi SP. Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. *J Proteome Res*. 2003; 2:43–50. [PubMed: 12643542]
4. Kall L, Storey JD, MacCoss MJ, Noble WS. Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *J Proteome Res*. 2008; 7:29–34. [PubMed: 18067246]
5. Choi H, Nesvizhskii AI. False discovery rates and related statistical concepts in mass spectrometry-based proteomics. *J Proteome Res*. 2008; 7:47–50. [PubMed: 18067251]
6. Navarro P, Vazquez J. A refined method to calculate false discovery rates for peptide identification using decoy databases. *J Proteome Res*. 2009; 8:1792–1796. [PubMed: 19714873]
7. Goloborodko AA, Mayerhofer C, Zubarev AR, Tarasova IA, Gorshkov AV, et al. Empirical approach to false discovery rate estimation in shotgun proteomics. *Rapid Commun Mass Spectrom*. 2010; 24:454–462. [PubMed: 20069687]
8. Lam H, Deutsch EW, Aebersold R. Artificial decoy spectral libraries for false discovery rate estimation in spectral library searching in proteomics. *J Proteome Res*. 2010; 9:605–610. [PubMed: 19916561]
9. Eng JK, McCormack AL, Yates IJR. An approach to correlate tandem mass spectral data of peptides with amino acid sequences. *J Am Soc Mass Spectrom*. 1994; 5:976–989.
10. Link AJ, Eng J, Schieltz DM, Carmack E, Mize GJ, et al. Direct analysis of protein complexes using mass spectrometry. *Nat Biotechnol*. 1999; 17:676–682. [PubMed: 10404161]
11. Washburn MP, Wolters D, Yates JR 3rd. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat Biotechnol*. 2001; 19:242–247. [PubMed: 11231557]
12. Nesvizhskii AI, Vitek O, Aebersold R. Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat Methods*. 2007; 4:787–797. [PubMed: 17901868]
13. Keller A, Nesvizhskii AI, Kolker E, Aebersold R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal Chem*. 2002; 74:5383–5392. [PubMed: 12403597]
14. Choi H, Nesvizhskii AI. Semisupervised model-based validation of peptide identifications in mass spectrometry-based proteomics. *J Proteome Res*. 2008; 7:254–265. [PubMed: 18159924]
15. Kall L, Canterbury JD, Weston J, Noble WS, MacCoss MJ. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat Methods*. 2007; 4:923–925. [PubMed: 17952086]
16. Andrews, S.; Tsochantaridis, I.; Hofmann, T. *Advances in Neural Information Processing Systems 15*. MIT Press; Vancouver, British Columbia: 2002. Support vector machines for multiple-instance learning; p. 561–568.
17. Bennett, KP. *Advances in Kernel Methods: Support Vector Learning*. MIT Press; Cambridge, MA: 1999. Combining support vector and mathematical programming methods for classification; p. 307–326.
18. Spivak M, Weston J, Bottou L, Kall L, Noble WS. Improvements to the percolator algorithm for Peptide identification from shotgun proteomics data sets. *J Proteome Res*. 2009; 8:3737–3745. [PubMed: 19385687]
19. Sanders SL, Jennings J, Canutescu A, Link AJ, Weil PA. Proteomics of the Eukaryotic Transcription Machinery: Identification of Proteins Associated with Components of Yeast TFIID by Multidimensional Mass Spectrometry. *Mol Cell Biol*. 2002; 22:4723–4738. [PubMed: 12052880]
20. Unnikrishnan A, Gafken PR, Tsukiyama T. Dynamic changes in histone acetylation regulate origins of DNA replication. *Nat Struct Mol Biol*. 2010; 17:430–437. [PubMed: 20228802]

21. Ross PL, Huang YN, Marchese JN, Williamson B, Parker K, et al. Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents. *Mol Cell Proteomics*. 2004; 3:1154–1169. [PubMed: 15385600]
22. Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, et al. *Saccharomyces Genome Database: the genomics resource of budding yeast*. *Nucleic Acids Res*. 2012; 40:D700–705. [PubMed: 22110037]
23. Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, et al. The Universal Protein Resource (UniProt). *Nucleic Acids Res*. 2005; 33:D154–159. [PubMed: 15608167]
24. Deutsch EW, Mendoza L, Shteynberg D, Farrah T, Lam H, et al. A guided tour of the Trans-Proteomic Pipeline. *Proteomics*. 2010; 10:1150–1159. [PubMed: 20101611]
25. McAfee KJ, Duncan DT, Assink M, Link AJ. Analyzing proteomes and protein function using graphical comparative analysis of tandem mass spectrometry results. *Mol Cell Proteomics*. 2006; 5:1497–1513. [PubMed: 16707483]
26. Niu, X.; McAfee, KJ.; Duncan, DT.; Assink, M.; Link, AJ. UT-ORNL-KBRIN Bioinformatics Summit 2008. *BMC Bioinformatics*; Cadiz, KY: 2008. A computational and analysis tool for proteomics research; p. 22
27. Washburn MP, Wolters D, Yates JR 3rd. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat Biotechnol*. 2001; 19:242–247. [PubMed: 11231557]
28. McDonald WH, Tabb DL, Sadygov RG, MacCoss MJ, Venable J, et al. MS1, MS2, and SQT-three unified, compact, and easily parsed file formats for the storage of shotgun proteomic spectra and identifications. *Rapid Commun Mass Spectrom*. 2004; 18:2162–2168. [PubMed: 15317041]
29. Chang CC, Lin CJ. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*. 2011; 2:1–27.
30. Bach, FR.; Lanckriet, GRG.; Jordan, MI. Fast kernel learning using sequential minimal optimization. Computer Science Division, University of California; Berkeley, CA: 2004.
31. Schwartz JC, Senko MW, Syka JE. A two-dimensional quadrupole ion trap mass spectrometer. *J Am Soc Mass Spectrom*. 2002; 13:659–669. [PubMed: 12056566]
32. Hu Q, Noll RJ, Li H, Makarov A, Hardman M, et al. The Orbitrap: a new mass spectrometer. *J Mass Spectrom*. 2005; 40:430–443. [PubMed: 15838939]
33. Makarov A, Denisov E, Kholomeev A, Balschun W, Lange O, et al. Performance evaluation of a hybrid linear ion trap/orbitrap mass spectrometer. *Anal Chem*. 2006; 78:2113–2120. [PubMed: 16579588]
34. Senko MW, Beu SC, McLafferty FW. Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. *JASMS*. 1995; 6:229–233.
35. Jiang X, Han G, Ye M, Zou H. Optimization of filtering criterion for SEQUEST database searching to improve proteome coverage in shotgun proteomics. *BMC Bioinformatics*. 2007; 8:323. [PubMed: 17761002]
36. Jones AR, Siepen JA, Hubbard SJ, Paton NW. Improving sensitivity in proteome studies by analysis of false discovery rates for multiple search engines. *Proteomics*. 2009; 9:1220–1229. [PubMed: 19253293]
37. Burges CJC. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*. 1998; 2:121–167.
38. Anderson DC, Li W, Payan DG, Noble WS. A new algorithm for the evaluation of shotgun peptide sequencing in proteomics: support vector machine classification of peptide MS/MS spectra and SEQUEST scores. *J Proteome Res*. 2003; 2:137–146. [PubMed: 12716127]
39. Fawcett T. An introduction to ROC analysis. *Pattern Recognition Letters*. 2006; 27:861–874.
40. Yeh ST. Using Trapezoidal Rule for the Area Under a Curve Calculation. 2002

Abbreviations

LC/MS/MS Liquid chromatography coupled with tandem mass spectrometry

PSM	peptide spectra match
FDR	false discovery rate
TPR	true positive rate
FPR	false positive rate
ROC	Receiver Operating Characteristic Curve
TP	true positive
TN	true negative
FP	false positive
FN	false negative
SVM	support vector machine
UPS1	Universal Proteomics Standard Set
DTT	dithiothreitol
IAA	iodoacetamide
RP	reverse-phase
nanoESI	nano-electrospray ionization
MS	mass spectrometry
MS/MS	tandem mass spectrometry
PBMC	peripheral blood mononuclear cells
PMN	polymorphonuclear cells
TEAB	triethylammonium bicarbonate
TFE	trifluoroethanol
MiPS	monoisotopic precursor selection
LC	liquid chromatography
SCX	strong cation exchange
TPP	Trans Proteomic Pipeline
MudPIT	multidimensional protein identification technology
D_n	total number of spectra matched to decoy peptide sequences
T_n	total number of the PSMs matched to target peptide sequence
AUC	Area Under the Curve
m+	number of target PSMs
mTM	number of decoy PSMs
S^+	the set of target PSMs
S^-	the set of decoy PSMs

Data: S^+, S^-, θ, γ
 Result: S^+
 Pre-process;
 Compute $x_{\bar{c}}$ and $d(x_i^+, x_{\bar{c}})$ for $i = 1, 2, \dots, m$;
 Sort $d(x_i^+, x_{\bar{c}})$ for $i = 1, 2, \dots, m$;
 Calculate the ratio of decoy PSMs to the total, and choose θ according to the ratio;
 Select $\theta|S^+|$ data points x_i^+ with smallest distance and remove them;
 Update S^+ ;
 Set $S = S^+ \cup S^-$;
 Refine process;
 Classification process;
 Train a SVM classifier f based on S ;
 Remove x_i^+ from S^+ if $f(x_i^+) \leq 0$;
 Update S^+ ;
 Set $S = S^+ \cup S^-$;
 Adjust process;
 Train a SVM classifier f based on S ;
 Sort $f(x_i^-)$ in descending order;
 Let $x_{\bar{\gamma}}$ be the $\gamma|S^-|$ th largest $f(x_i^-)$;
 Remove x_i^+ from S^+ if $f(x_i^+) \leq f(x_{\bar{\gamma}}^-)$;
 Post-process;
 Keep all full digested PSMs;
 Keep the top τ_{half} half digested PSMs according to PSMevaluator;
 Keep the top τ_{non} non-digested PSMs according to PSMevaluator;

Notation Summary

x_i^+ the i th +1 PSM (target PSM)
 x_i^- the i th -1 PSM (decoy PSM)
 $x_{\bar{c}}$ the centroid of all -1 PSMs in the feature space
 $d(x_i^+, x_{\bar{c}})$ the distance between x_i^+ and $x_{\bar{c}}$ in the feature space
 θ the ratio of assumed false positive PSMs to total PSMs
 m total number of PSMs
 S the set of m PSMs
 S^+ the set of +1 PSMs
 S^- the set of -1 PSMs
 S_0 the union of S^- and S^+ updated after the first round
 S_1 the set of PSMs obtained by removing x_i^+ from S_0 if $f(x_i^+) \leq 0$
 γ the ratio of number of retained decoy PSMs in S_1 to the total number of decoy PSMs in S_1
 S_2 the set of PSMs obtained by removing x_i from S_1 if $f(x_i) \leq f(x_{\bar{\gamma}}^-)$
 τ_{half} the ratio of the number of half-canonical PSMs accepted to the total number of half-canonical PSMs in S_2
 τ_{non} the ratio of the number of non-canonical PSMs accepted to the total number of non-canonical PSMs in S_2

Figure 1.
 Pseudocode for the De-Noise algorithm.

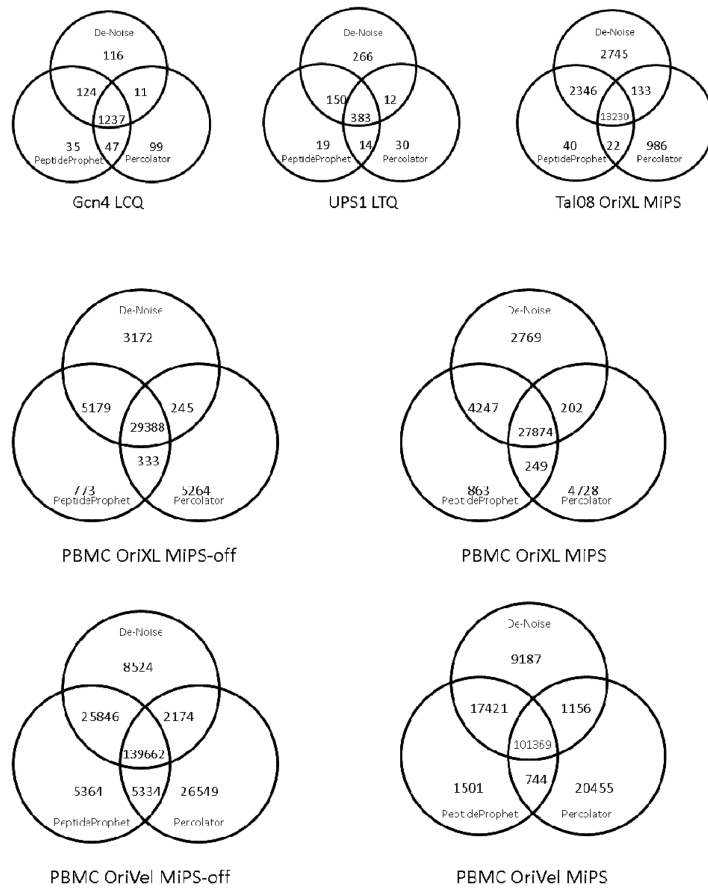
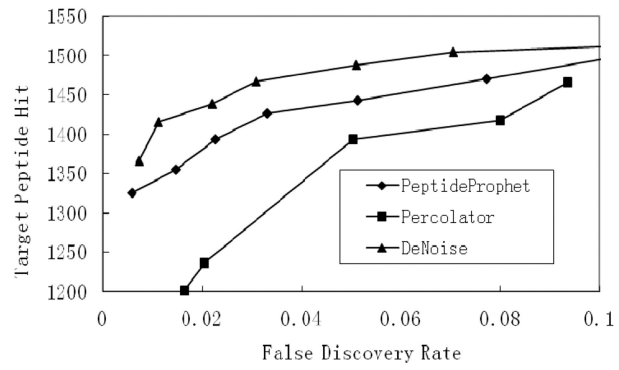
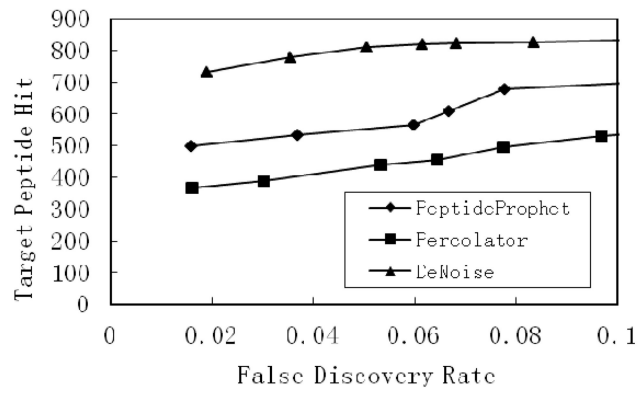


Figure 2. Venn diagrams for the seven datasets showing the number of overlapping validated PSMs from De-Noise, PeptideProphet, and Percolator. An FDR of 0.05 was used for all three approaches.

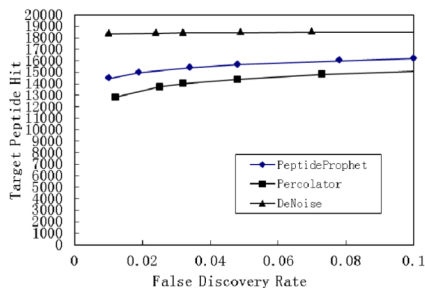
A. Gcn4 LCQ



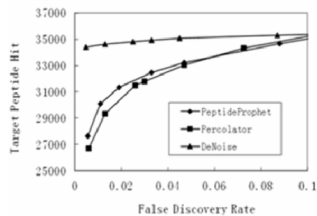
B. UPS1 LTQ



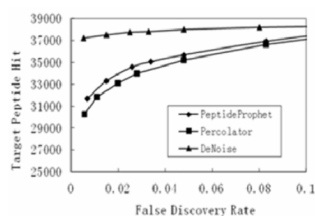
C. Tal08 OriXL MiPS



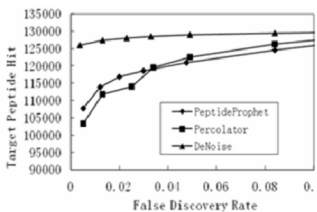
D. PBMC OriXL MiPS-off



E. PBMC OriXL MiPS



F. PBMC OriVel MiPS



G. PBMC OriVel MiPS-off

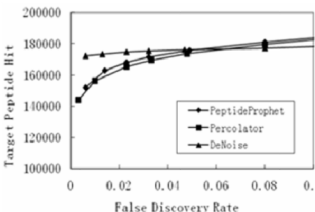
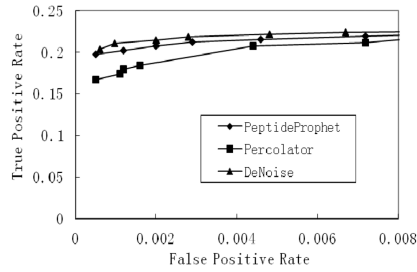


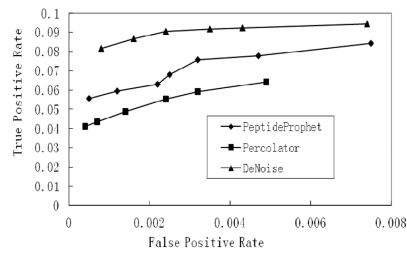
Figure 3.

Plots of target PSM hits for the seven datasets validated under a series of FDRs for DeNoise, PeptideProphet, and Percolator. The number of target peptide hits is plotted for a FDR range from 0.01 to 0.1. (A) Gcn4 LCQ (B) UPS1 LTQ (C) Tal08 LTQ-Orbitrap XL MiPS (D) PBMC LTQ-Orbitrap XL MiPS (E) PBMC LTQ-Orbitrap XL MiPS-off (F) PBMC LTQ-Orbitrap Velos MiPS (G) PBMC LTQ-Orbitrap Velos MiPS-off.

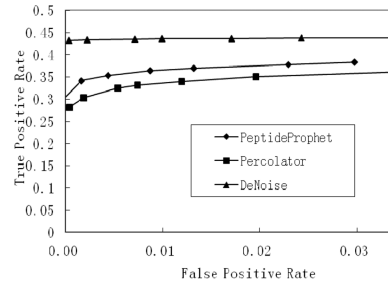
A. Gcn4 LCQ



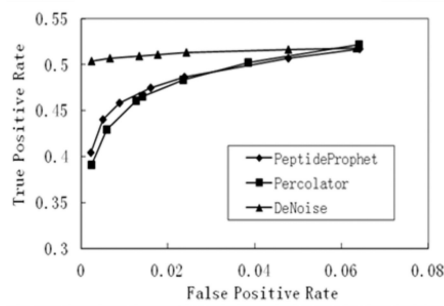
B. UPS1 LTQ



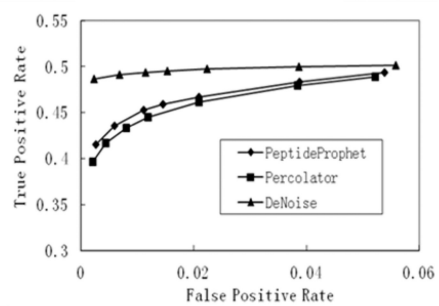
C. Tal08 OriXL MiPS



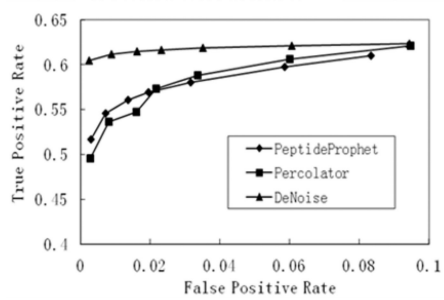
D. PBMC OriXL MiPS-off



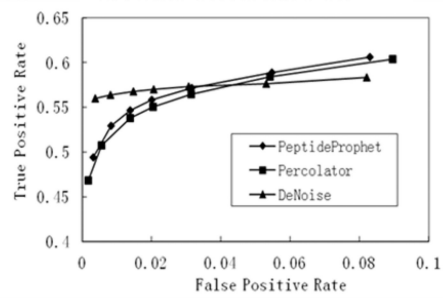
E. PBMC OriXL MiPS



F. PBMC OriVel MiPS



G. PBMC OriVel MiPS-off



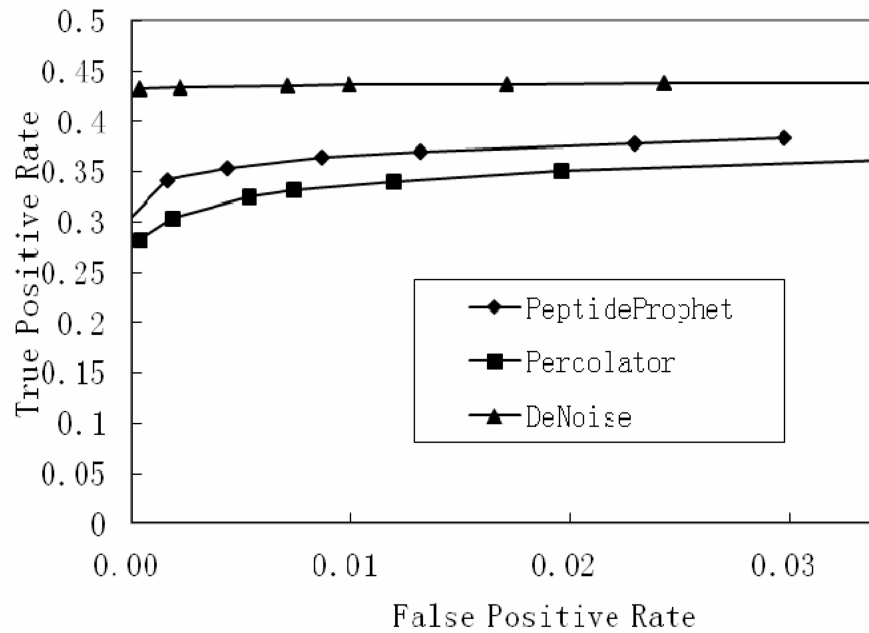


Figure 4. ROC curves for the seven datasets showing the validation performance of De-Noise, PeptideProphet, and Percolator. (A) Gcn4 LCQ (B) UPS1 LTQ (C) Tal08 LTQ-Orbitrap XL MiPS-off (D) PBMC LTQ Orbitrap XL MiPS (E) PBMC LTQ-Orbitrap XL MiPS-off (F) PBMC LTQ-Orbitrap Velos MiPS (G) PBMC LTQ-Orbitrap Velos MiPS-off.

Table 1

Summary of LC-MS/MS Dataset and SEQUEST Search Result

Sample	Mass Spectrometer	MIPS	PSM	Target PSM	Decoy PSM	FDR	Decoy PSMs/Total PSMs
Gen4	LCQ	N/A	14892	6703	8189	1.09	0.55
UPS1	LTD	N/A	17335	8974	8361	0.96	0.48
Tal08	Orbitrap XL	ON	69560	42222	27338	0.79	0.39
PBMC	Orbitrap XL	ON	103679	68334	35345	0.64	0.34
PBMC	Orbitrap XL	OFF	117751	76395	41356	0.70	0.35
PBMC	Orbitrap Velos	ON	301879	208765	93114	0.62	0.31
PBMC	Orbitrap Velos	OFF	447350	307549	139801	0.62	0.31
PMN	LTD	N/A	51423	29992	21431	0.83	0.42

Note: PSM=Peptide Spectrum Match

Table 2

Summary of Unfiltered, Categorized PSMs

Dataset	Target			Decoy			Decoy/Target Ratio		
	Fully-Tryptic	Half-Tryptic	Non-Tryptic	Fully-Tryptic	Half-Tryptic	Non-Tryptic	Fully-Tryptic	Half-Tryptic	Non-Tryptic
Gcn4LCQ	1453	1210	4040	106	1465	6618	0.07	1.211	1.638
UPSIL1TQ	645	2013	6316	236	2588	5537	0.36	1.286	0.877
Tal08 OriXL MiPS	14893	6809	20520	419	5877	21042	0.028	0.863	1.025
PBMC OriXL MiPS	26760	15647	25927	737	8583	26025	0.03	0.548	1.004
PBMC OriXL MiPS-off	28561	17490	30344	948	10333	30075	0.03	0.590	0.991
PBMC OriVel MiPS	110404	35915	62446	2520	24682	65912	0.023	0.687	1.056
PBMC OriVel MIS-off	134117	77052	96380	3414	34985	101402	0.025	0.454	1.052
PMN LTQ	8257	5946	15789	376	4752	16303	0.046	0.799	1.033

Note: Target=Target PSMs, Decoy=Decoy PSMs

Table 3

Summary of PSMs Validated by Different Approaches

Approach	Gcn4 LCQ		UPSI LTQ		Tai08 OriXL MIPS		PMN LTQ	
	Target PSMs	Decoy PSMs	Target PSMs	Decoy PSMs	Target PSMs	Decoy PSMs	Target PSMs	Decoy PSMs
SEQUEST	6703	8189	8974	8361	42222	27338	29992	21431
PeptideProphet	1443	38	566	18	15638	387	8957	220
Percolator De-Noise	1394 1488	35 39	438 811	12 21	14371 18454	354 475	9060 11341	219 287
Low-Stringency	1128	18	293	6	9979	108	11252	553
High-Stringency	588	11	154	0	7083	272	8411	134
Approach	PBMC OriXL MIPS		PBMC OriXL MIPS-off		PBMC OriVel MIPS		PBMC OriVel MIPS	
	Target PSMs	Decoy PSMs	Target PSMs	Decoy PSMs	Target PSMs	Decoy PSMs	Target PSMs	Decoy PSMs
SEQUEST	68334	33545	76395	41356	208372	93112	307459	139081
PeptideProphet	33233	802	35673	869	120961	2947	175790	4393
Percolator	33053	793	35230	866	122568	3133	173719	4363
De-Noise	35070	813	37894	927	128977	3272	176206	4287
Low-Stringency	6135	235	25761	346	107693	1372	127857	1728
High-Stringency	11768	335	20963	930	71876	6539	120180	9043

Table 4

Summary of Overlapping, Validated PSMs

Dataset	PSMs Shared between Peptide and De-Noise	% PeptideProphet shared by De-Noise	PSMs Shared between Percolator and De-Noise	% Percolator Shared by De-Noise	PSMs Shared between Percolator and PeptideProphet	% Percolator Shared by PeptideProphet
Gen4 LCQ	1361	94	1248	90	1284	92
UPS1 LTQ	533	94	395	90	397	90
Tal08 OriXL-MiPS	15576	99	12763	89	13252	92
PBMC OriXL-MiPS	32121	97	28076	85	28123	85
PBMC OriXL MiPS-off	34567	97	29633	84	29721	84
PBMC OriVel MiPS	118790	98	102525	84	102113	83
PBMC OriVel MiPS-off	165508	94	141836	82	144996	83

Note: Data used in this table are outputs under FDR=0.05

Table 5

Summary of Overlapping Validated PSMs

Dataset	# of Shared PSMs between PeptideProphet and De-Noise			# of Shared PSMs between Percolator and De-Noise			# of Shared PSMs between PeptideProphet and Percolator		
	Full	Half	Non	Full	Half	Non	Full	Half	Non
Gcn4 LCQ	1314	47	0	1206	42	0	1237	46	1
UPS1 LTQ	399	125	9	276	112	7	270	115	12
Tal08 OriXL MiPS	14528	1041	7	12881	482	0	12767	485	0
PBMC OriXL MiPS	25962	6046	113	23565	4423	88	23431	4566	126
PBMC OriXL MiPS-off	27604	6802	161	24420	5070	143	24247	5204	270
PBMC OriVel MiPS	107595	11149	46	94718	7771	36	93830	8171	112
PBMC OriVelMiPS-off	130259	34738	511	115919	25735	182	114134	29234	1628

Note: Data used in this table are outputs under FDR=0.05

Table 6

Distribution of Validated PSMs

Dataset	De-Noise			PeptideProphet			Percolator		
	Full	Half	Non	Full	Half	Non	Full	Half	Non
Gen4 LCQ	1370	106	12	1375	68	1	1342	51	1
UPS1 L1TQ	597	190	24	403	147	16	278	144	17
Tal08 OriXL MiPS	14860	3471	123	14539	1088	11	13855	516	0
PBMC OriXL MiPS	26699	8228	165	25977	7006	250	27025	5865	163
PBMC OriXL MiPS-off	28498	9292	194	27622	7649	402	28271	6661	298
PBMC OriVel MiPS	110138	18490	349	107730	13001	230	11990	10453	125
PBMC OriVelMiPS-off	133846	41752	608	130321	42835	2633	133436	38069	2214