Computational Strategies for Proteogenomics Analyses

by

Andy Kong

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in The University of Michigan
2017

Doctoral Committee:

Associate Professor Alexey I. Nesvizhskii, Chair
Professor Philip C. Andrews
Assistant Professor Yuanfang Guan
Assistant Professor Ryan E. Mills
Associate Professor Maureen Sartor

Andy T. Kong

andykong@umich.edu

ORCID ID: 0000-0002-4708-7815

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

ABSTRACT

Proteogenomics is an area of proteomics concerning the detection of novel peptides and peptide variants nominated by genomics and transcriptomics experiments. While the term primarily refers to studies utilizing a customized protein database derived from select sequencing experiments, proteogenomics methods can also be applied in the quest for identifying previously unobserved, or missing, proteins in a reference protein database. The identification of novel peptides is difficult and results can be dominated by false positives if conventional computational and statistical approaches for shotgun proteomics are directly applied without consideration of the challenges involved in proteogenomics analyses. In this dissertation, I systematically distill the sources of false positives in peptide identification and present potential remedies, including computational strategies that are necessary to make these approaches feasible for large datasets.

In the first part, I analyze high scoring decoys, which are false identifications with high assigned confidences, using multiple peptide identification strategies to understand how they are generated and develop strategies for reducing false positives. I also demonstrate that modified peptides can cause violations in the target-decoy assumptions, which is a cornerstone for error rate estimation in shotgun proteomics, leading to potential underestimation in the number of false positives. Second, I address computational bottlenecks in proteogenomics workflows through the development of two database search engines: EGADS and MSFragger. EGADS aims to address issues relating to the large sequence space involved in proteogenomics studies by using graphical processing units to accelerate both in-silico digestion and similarity scoring. MSFragger implements a novel fragment ion index and searching algorithm that vastly speeds up spectra similarity calculations. For the identification of modified peptides using the open search strategy, MSFragger is over 150X faster than conventional database search tools. Finally, I will discuss refinements to the open search strategy for detecting modified peptides and tools for improved collation and annotation. Using the speed afforded by MSFragger, I perform open searching on

several large-scale proteomics experiments, identifying modified peptides on an unprecedented scale and demonstrating its utility in diverse proteomics applications.

The ability to rapidly and comprehensively identify modified peptides allows for the reduction of false positives in proteogenomics. It also has implications in discovery proteomics by allowing for the detection of both common and rare (including novel) biological modifications that are often not considered in large scale proteomics experiments. The ability to account for all chemically modified peptides may also improve protein abundance estimates in quantitative proteomics.

CHAPTER I

INTRODUCTION TO SHOTGUN PROTEOMICS AND PROTEOGENOMICS

1.1 Venturing beyond the reference proteome

In May 2014, more than thirteen years after the draft of the human genome [1,2], two studies from independent groups appeared in Nature each claiming to have completed the first draft of the human proteome [3,4]. These drafts of the human proteome were based on mass spectrometry-based proteomics, the now dominant tool for large-scale proteome analysis. In these studies, data was generated from diverse human tissues, or otherwise aggregated from public repositories, to build a comprehensive catalog of human proteins. From the roughly 20,000 protein-coding genes, both studies reported mass-spectrometry evidence for over 17,000 protein-coding genes. Immediately after these heavily publicized publications, much criticism was raised in the proteomics community regarding the lax false discovery rate (FDR) filtering used in the studies. Several re-analyses of these datasets only identified 13-14,000 proteins [5] and others raised concerns about the large number of olfactory receptor proteins identified in non-nasal tissues (allegedly due to poor quality spectra and non-unique peptides) [6]. As a result, guidelines were established for calling novel (as was the case for many peptides and proteins in these studies) identifications [7] and subsequent work from one of the groups publishing the draft proteome have revised downwards their observed number of proteins to less than 15,000 [8], more than 2,000 fewer than what was originally claimed. The drafts of the human proteome and the subsequent controversy illustrate the numerous statistical pitfalls and computational challenges surrounding mass spectrometry-based proteomics.

The concept of novel identifications in attempts to define the proteome can be a bit misleading as modern high-throughput proteomics studies relies on databases of protein sequences derived from genomics and transcriptomics studies. Hence, defining the proteome is more of an exercise in validating predicted protein products rather than directly observing proteins and their sequences. This is in contrast to the efforts in the early 1990s, prior to the availability of genomic or protein databases, where Edman sequencing of intact proteins or enzymatically-digested fragments was used to generate partial protein sequences for which degenerate oligonucleotide primers could be made to PCR the gene for sequencing [9]. Presently, the majority of proteomics studies utilize high quality reference databases generated from a combination of manual and automated curation such as UniprotKB [10] and Ensembl [11]. Others use custom protein databases [12] derived from related sequencing experiments to capture variants (both point and splice) [13] and non-canonical regions of coding potential (such as lncRNAs [14]). Studies of the latter type have been classified under the label of proteogenomics [7] in recent years as an emerging field of proteomics (Figure 1-1). Despite this classification, the methods and challenges involved in identifying novel peptides and proteins, whether they stem from entries in the reference database that have not been previously observed by mass spectrometry [5,15] or from predictions from sequencing data, are largely the same (rather unsurprising as the reference protein database is also derived from sequencing efforts). These efforts to characterizing the novel and unknown may also extend to the vast repertoire of post-translational modifications (PTMs) that regulate most proteins [16]. Together, they embody the frontiers of the observable proteome. In this dissertation, we define some of the challenges in these explorations beyond the reference proteome and present computational tools and strategies that address these issues.

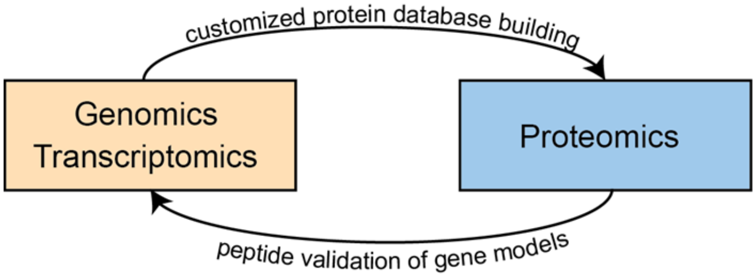

customized protein database building
Genomics
Transcriptomics
Proteomics
peptide validation of gene models

Figure 1-1 Proteogenomics workflow.
Genomics and transcriptomics data are used to generate customized protein databases that are used for peptide identification. The identified peptides are then used to validate or refine gene models.

1.2 Mass spectrometry-based proteomics

Modern high-throughput shotgun proteomics has its origins in the development of four crucial technologies in the 1980s and early 1990s. The first are the advances in peptide and protein ionization technologies such as MALDI (matrix-assisted laser desorption/ionization) and electrospray ionization that enabled the analysis of these biomolecules in mass spectrometers [17]. In particular, electrospray ionization is compatible with high performance liquid chromatography (HPLC) systems that are critical for separating peptides in complex biological mixtures. Second, the development of tandem (MS/MS) mass spectrometry enabled the rapid sequencing of peptides bypassing lengthy Edman sequencing [18]. In tandem mass spectrometry, peptide ions are collided or otherwise fragmented into peptide fragments and the mass spectra of the resultant fragments are collected. The peptides fragment primarily at the amide bonds between residues, allowing the sequence to be deduced by comparing the mass differences between series of ions to the mass of amino acids. The third are automated instrument control methods that enable mass spectrometers to dynamically select which ions to fragment based on the ions it observes eluting from chromatographic system in real time [19]. These data-dependent acquisition methods enable mass spectrometers to focus on the ions most likely to yield additional information about the sample (e.g. high intensity ions that have not already been fragmented, ions that do not correspond to signals of known chemical noise) as the number of eluting ions is far greater than the rate at which mass spectrometers can acquire tandem mass spectra. Finally, and of particular focus in this dissertation, computational tools that can interpret and assign a peptide identification to these tandem mass spectra. Without these automated tools, the labor involved in manual interpretation would prohibit mass spectrometry-based proteomics from evolving into a high-throughput technology.

In practice, the modern realizations of these technologies in shotgun proteomics are as follows [20]. Proteins are enzymatically digested, commonly with trypsin, into peptides that are loaded into an HPLC system coupled to a mass spectrometer. To reduce sample complexity (allowing lower abundance ions to be sampled in the mass spectrometer – resulting in greater proteome coverage), proteins and/or peptides can be fractionated in gels or another liquid chromatography system prior to loading. As peptides are eluted from the HPLC system, they are ionized and the mass spectrometer generates a mass spectrum containing all eluting peptide ions (known as a survey or MS1 scan). Based on the acquired survey scan, the onboard computer selects a number of peptide ions to select for fragmentation (often the most abundant peptide ions). The mass spectrometer then iteratively applies a mass filter to the incoming stream of ions, selecting a narrow mass window around the selected peptide ion, and fragments the peptide ion into fragments. The fragment ions are then detected and recorded as a tandem (also called MS/MS, MS2, or fragmentation) spectrum. After a number of tandem mass spectra have been acquired, the selected masses are placed on an dynamic exclusion list [21] to avoid collecting additional mass spectra of the same peptide ion and a fresh survey scan is collected to select the next targets for fragmentation. The collected mass spectra can then be used for peptide identification and quantitation.

While this model of data dependent acquisition has served as the primary workhorse of shotgun proteomics for nearly two decades, advances in instrument speed and accuracy in recent years have led to dramatic increases in the depth of proteome coverage and rate of acquisition. Detection of over 10,000 proteins is now routine for fractionated human cell line data [22,23] while single shot analysis of the yeast proteome (approximately 4000 proteins) can now be performed in an hour [24], in contrast to the 144 hours of analysis time required for experiments performed as recently as 2008 [25]. As a result of the ever-increasing instrument speed, the amount of experimental data generated has increased significantly. This is readily observed through the growth in the amount of data stored in public repositories of proteomics experiments (Figure 1-2).

Much of this growth appears to be stimulated by the introduction of instruments that are able to acquire tens of high-resolution (tens of parts per million or less) tandem mass spectra per second such as the TripleTOF 5600 [26] or the Q Exactive [27] mass spectrometers. A deeper inspection of the PRIDE repository [28] (Table 1-1) reveals that the majority of deposited data is indeed generated from these and related instruments.

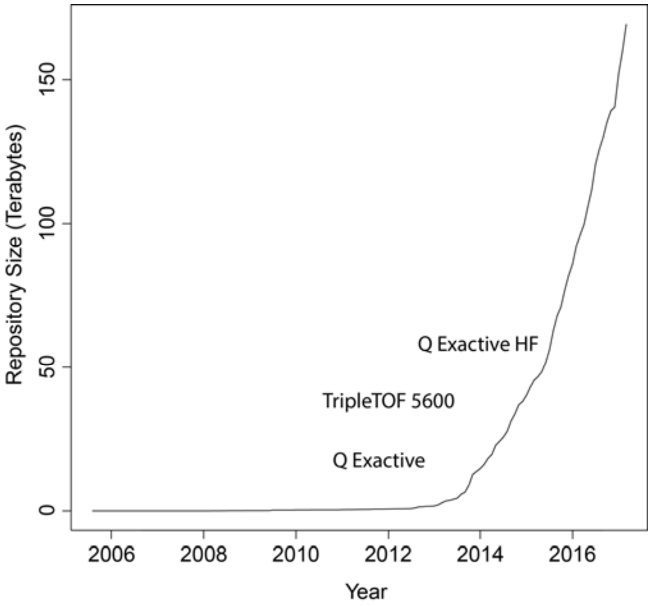


Figure 1-2 Size of the PRIDE data repository of proteomics data over time as of March 2017.
Project sizes are determined from the contents of their FTP directories. Project publication dates are used rather than submission dates. Popular mass spectrometers are plotted near their approximate release dates.

Table 1-1 Top mass spectrometers in PRIDE as of March 2017.
Instruments are extracted from PRIDE project pages with sizes determined from projects' FTP directories. Projects analyzed using multiple instrument types are associated with all instruments, leading to some duplicate counting.

Instrument	Projects	Total Size (TB)	Year of Introduction
Q Exactive	740	72.52	2011 [27]
LTQ Orbitrap Velos	770	51.07	2009 [29]
LTQ Orbitrap*	600	28.49	2005 [30]
LTQ Orbitrap Elite	270	16.61	2012 [31]
TripleTOF 5600	191	10.08	2011 [26]

*In many projects, the annotations on the PRIDE project pages are incorrect, with experiments performed on LTQ Orbitrap Elite and LTQ Orbitrap Velos incorrectly labeled as LTQ Orbitrap.

1.3 Methods for peptide identification in shotgun proteomics

The automated interpretation of tandem mass spectra is an important part of any shotgun proteomics workflow. Peptide identification algorithms are assessed on both their ability to correctly identify peptides as well as their computational runtime as peptide identification is often a computational bottleneck in many proteomics workflows [20]. These runtime concerns may be more relevant than ever due to the volume of data generated by modern high-speed instruments. As previously described, peptide ions are typically fragmented at the amide bonds between residues in tandem mass spectrometry. However, the fragmentation does not occur uniformly [32] and the fragmentation spectrum often fail to contain a complete ion ladder for easy interpretation. Hence, there are multiple strategies for interpreting tandem mass spectra.

There are three main classes of tools for peptide identification: database searching, tag-based searching, and de novo sequencing. The lines between the classes are often blurred as ideas between the classes can be combined to form hybrid strategies. One of the earliest tools to gain widespread adaptation was the database search tool SEQUEST [33] (also re-implemented as the open-source tool Comet [34]), which interpreted experimental spectrum by computing a cross correlation function against theoretical spectra derived from in-silico digested peptides in a protein database. This database search model has been adapted by numerous other tools including Mascot [35], X! Tandem [36], Andromeda [37], MS-GF+ [38] and countless others – varying in how they score or compare the experimental spectrum against theoretical spectra, or how the search space is partitioned and prioritized. Due to their performance in both accuracy and speed, they remain the dominant class of tools for peptide identifications. As more and more fragmentation spectra are collected and assembled in public repositories, the use of spectral libraries for peptide identifications have also emerged, despite initial comments that there are simply too many peptides to build an effective spectral library [33]. In spectral library searching, previously observed and identified fragmentation spectra are assembled [39] and compared against experimental spectra. The known fragmentation patterns (intensities of fragment ions) can be more discriminatory than the theoretical spectra predicted by database search tools but the approach is limited only to the previously identified peptides present in the spectral library.

While the complete ion ladder is often missing from fragmentation spectra, there is often a partial ladder that reveals a subsequence of the peptide. Tag-based approaches use this information by looking for a partial ladder and using the derived information to filter a sequence database [40]. Compared to database searching, the tag-filtered list of candidates is much smaller so more computational intensive operations, such as the identification of unknown modifications [41], can be performed. De novo sequencing tools attempt to identify peptides from fragmentation spectra without the use of a reference database [42,43]. The main benefit of de novo sequencing tools is that they allow for the identification of peptides that are not known or present in the reference database. They can also be used to identify post-translational modifications [44]. However, they are computationally expensive to use and require high-quality spectra, making them impractical in most high-throughput experiments. Hybrid de novo sequencing methods have been developed to use information from a reference database [45] or fragmentation information from spectral libraries [46].

Regardless of the approach used to identify peptides in a fragmentation spectrum, the output of each tool is the identified peptide, a raw score representing the quality or confidence of the assignment based on the tool's internal scoring mechanism, and a calibrated score (such as a probability or expectation value) that normalizes the raw score (as it can be heavily dependent on peptide length or spectrum complexity) allowing for quality comparisons between different peptide to spectrum matches (PSMs). However, the quality and accuracy of these calibrated scores is much debated [47].

1.4 Error rate estimation in shotgun proteomics

While peptide identification tools provide scores that estimates the quality of the peptide spectrum match, they do not provide any estimates on the number of false identifications in a given experiment. This problem is complicated by the fact that scores are not comparable between different identification tools and that the quality assessments of individual PSMs do not have access to information that can be derived from experiment-wide observations, such as the frequency of missed cleavages or instrument mass accuracy and calibration. Hence, peptide validation methods were developed that combines peptide identification scores with auxiliary information in a statistical framework that can estimate the number of correct and incorrect hits. One of the pioneering tools for peptide validation is PeptideProphet [48] which computed a discriminant score for each PSM and used the expectation-maximization algorithm to model the distributions of correct and incorrect hits. The models can then be used to estimate the probability of correct hits and establish thresholds for controlling error rates in an experiment. These principles also have been extended to proteins to control for error rates at the protein level [49].

Complementary to these modeling based methods, the use of decoy sequences is widespread in shotgun proteomics [50]. Decoy sequences are peptide or protein sequences that are artificial and should not be found in nature. Hence, by searching against decoy sequences, it is possible to estimate the rate of error matches or the distribution of random scores. Decoy sequences were first generated from reversed protein sequences [51] but reversed peptides have been shown to perform similarly and may be superior for high mass accuracy data as they preserve the distribution of peptide masses. Decoys have also been incorporated into peptide and protein validation models as negative examples in semi-supervised modeling [52,53]. In addition to serving as negative examples in statistical modeling, they can also be used to directly and empirically estimate the false discovery rate in a given experiment. In a common target-decoy approach, a combined database is generated with targets (the protein database of interest) and decoys of equal length (from reversed proteins or peptides). Experimental fragmentation spectra are then searched against this combined database using a peptide identification tool. Under the target-decoy assumption that incorrect assignments will match equally to target and decoy sequences, the number of false positives is equal to the number of decoy matches so the false discovery rate can be estimated by dividing the number of decoy hits to the number of target hits. A scoring threshold can then be set to achieve a target false discovery rate. Not only does the target-decoy approach allow estimation of false discovery rates at the PSM level, it can also be used to estimate FDR at the peptide or protein level.

Error rate estimation in shotgun proteomics remains an area of active development. This is especially true for error rate estimation in very large datasets [8,54,55] and the use of different prior probabilities by integrating abundance information from transcriptomics or proteomics repositories [55,56], both of which are relevant in a proteogenomics context.

1.5 False positives in proteogenomics analyses

While false positives are produced as part of any shotgun proteomics analysis workflow, there are two key considerations that are of particular importance for proteogenomics: class-specific FDR estimation and false identifications of nonrandom nature [7]. In conventional proteomics analyses, a score threshold is selected to achieve a particular FDR for the entire dataset, typically using the target-decoy approach. This approach has proved problematic for proteogenomics analyses as the likelihoods for identifying novel peptides are different than for abundant peptides and much stronger evidence is needed to support a novel identification. Failure to account for these differences have cause dramatic underestimation of the FDR for novel peptides [57]. In one proteogenomics study examining aberrant peptides in cancer, a global 1% FDR filter resulted in a 36% FDR amongst the novel peptide identifications [58]. Performing FDR estimation separately for known peptides and novel peptides allowed a 1% FDR to be achieved in both classes. A theoretical analysis of this problem has demonstrated that a more complete annotation of the genome is linked to the degree of FDR underestimation for novel peptides (when a global FDR filter is used) [59]. While there are proteogenomics studies that employ class specific FDR estimation [60], the vast majority do not [3,4,13], leading to over-reporting of novel discoveries.

The second major challenge facing proteogenomics analyses is the high degree of similarity between certain novel candidates and previously identified peptides in the reference database. Many novel peptides in proteogenomics originate from non-synonymous single nucleotide polymorphisms in the sequencing data. As these single amino acid substitutions can be similar or identical in mass with common chemical modifications, spectra of modified peptides can be misidentified as a novel variant peptide [61]. For example, the change in mass for an alanine to serine substitution or a phenylalanine to tyrosine substitution is identical to that of oxidation, an abundant chemical modification. As these incorrect assignments are due to chemically modified peptides, their fragmentation patterns are of a non-random nature and these errors may not be well modeled by the target-decoy strategy. While a number of such errors have been manually curated from the human draft proteome studies [5], there is a need to systematically annotate and study such errors.

1.6 Outline

False positives present two major challenges for proteogenomics analyses. The first concerns the sensitivity of proteogenomics experiments. Discovery of novel peptides involves the testing of large hypothesis spaces generated from sequencing experiments, requiring much stronger evidence for the identification of novel peptides when class specific FDR estimation is applied. When there is an abundance of false positives, the sensitivity of the experiment can be greatly affected making it impossible to detect any novel peptide at a reasonable FDR. The second involves chemically or biologically modified peptides that might be misidentified as a novel variant peptide. It is unclear whether these false positives are correctly modeled by the target-decoy approach as spectra from modified peptides can resemble spectra of novel variant peptides or other peptides in the reference database and may not match with equal propensity to target and decoy sequences. Even if these misidentifications are effectively captured by the target-decoy strategy, their presence as false positives in proteogenomics analyses might also decrease sensitivity.

The overall aim of this dissertation is to develop computational strategies that address these concerns in proteogenomics analyses. In Chapter Two, I examine the causes and composition of false positives in shotgun proteomics while exploring the effects of modified peptides in false discovery rate estimation using multiple peptide identification tools. In Chapter Three, I develop two database search tools that reduce the computation time required in proteogenomics analyses, enabling more comprehensive peptide identifications that can reduce false positives. In Chapter Four, I refine the open search strategy for identifying modified peptides and apply it to large scale proteomics experiments for profiling modified peptides in a number of proteomics experiments and applications.

CHAPTER II

CHARACTERIZATION AND REMEDIATION OF FALSE POSITIVES IN PROTEOGENOMICS WORKFLOWS

Portions of this chapter comparing narrow window and open search results have been published in Nature Methods [62]

2.1 Introduction

Manual validation of database search results have revealed several common sources of false positives including poor quality spectra, non-enzymatic cleavage, chemical or post-translational modifications, incorrect monoisotopic peak assignment, and incorrect charge state assignment [63,64]. While computational strategies have been developed to address a number of these sources (semi-tryptic searches, variable modification searches with common chemical modifications, isotope error correction in database search engines etc.), they are not comprehensive and are not always applied for large datasets due to the computational costs involved. In proteogenomics contexts where the presence of false positives is of much greater concern, these additional computations might be necessary for successful identification of novel peptides. We developed a computational framework for annotating and quantifying the false positives that might be avoided using more extensive analyses. Finding a large number of high confidence false positives is inherently difficult as they are indistinguishable from true positives on the basis of database search and peptide validation scores. Fortunately, the use of the target-decoy strategy [50] provides us with a mechanism to study false positives as decoy assignments are incorrect by design. We performed peptide identification using three database search engines [34,36,38] and a blind modification search tool [41] in both tryptic and semi-tryptic modes to quantify the false positives that are due to ambiguous assignments, semi-enzymatic cleavage, and modified peptides.

Next, we took a deeper look at modified peptides to see how they cause false positives and whether such false positives are correctly modeled by the target-decoy strategy. While blind modification search tools can identify modified peptides with unknown modifications, they suffer from reduced sensitivity and are often incompatible with other tools in common proteomics workflows such as those for peptide validation. Hence, there was considerable interest in a recent report [65] exploring the feasibility of searches using wide precursor mass tolerances of hundreds of Daltons (open searches) to identify modified peptides using conventional database search tools. We applied the open searching concept to identify modified peptides and compared the identifications with those from a conventional narrow window search.

2.2 Materials and methods

2.2.1 Characteristics and sources of high scoring decoys

Datasets and data preparation

A publicly available dataset consisting of a panel of various triple negative breast cancer cell lines and tissue specimens analyzed on a Thermo Scientific Q Exactive mass spectrometer [66] was downloaded from ProteomicsDB (PRDB004167) in vendor .raw format. The Thermo .raw files were converted to the mzML format using vendor provided centroiding and default parameters using the msconvert.exe tool from ProteoWizard (3.0.7398 64-bit version). For the MODa analysis, the mzML files were further converted to the MGF file format from the mzML files using default parameters.

A human protein database was retrieved from UniprotKB (download date: 2016-07-29) and appended with decoy proteins containing reversed peptide sequences (with prolines left in-place when they are immediately before a trypsin cleavage site to ensure that the distribution of tryptic peptide masses is identical between the forward and decoy space). Common contaminants (cRAP protein sequences from GPMDB and contaminants from MaxQuant) were appended to the concatenated protein database.

Peptide identification pipeline

Peptide identification was performed independently using three different database search engines (Comet[34] 2015.02 rev 1, X! Tandem[36] 2015.04.01.1, MS-GF+[38] v10089) and a blind-modification search tool (MODa[41] v1.51) against the protein database described above in both fully tryptic and semi-tryptic modes. For all searches, trypsin was specified as the enzyme used for digestion with the precursor mass tolerance set to 20ppm. Static carboxyamidomethylation (+57.021464 Da) on cysteine was specified for all tools while oxidation (+15.9949 Da) on methionine and N-terminal acetylation (+42.0106 Da) were specified as variable modifications in the three database search engines. For Comet and X! Tandem, up to five missed cleavages were allowed (MS-GF+ does not limit the number of missed cleavages). Parent isotopic error correction was enabled in all search engines (isotope_error = 1 in Comet and –ti "0,2" in MS-GF+). In Comet, the use of neutral loss was disabled and high-resolution MS/MS settings (fragment bin offset 0.0 and fragment bin tolerance of 0.02) were used for scoring. In X! Tandem, the top 100 peaks were used for scoring with a required minimum of 4 matched fragment peaks. For MODa analysis, only a single modification was permitted per peptide with a modification range of -500 to +500 Da. Fragment mass tolerance was set to 0.02 and the high-resolution MS/MS mode was enabled.

X! Tandem output files were converted to the pepXML format using the Tandem2XML tool found in the Trans-Proteomics Pipeline[67] (TPP) version 4.8.0. MS-GF+ outputs were converted from mzIdentML to pepXML using idconvert from ProteoWizard (3.0.6002).

Peptide validation and false discovery rate estimation

Peptide validation was performed individually on each pepXML file in the six (three search engines; tryptic and semi-tryptic) sets of outputs using PeptideProphet[48] (TPP 4.8.0) using the following settings: 'd' (report decoy hits), 'A' (high mass accuracy model), 'E' (use search engine calculated expectation values), 'P' (semi-parametric modeling), 'PPM' (parts per million in mass model). For each set of analyses, the pepXML files corresponding to the 457 LC-MS/MS runs were read using a custom Java program and false discovery rate estimation was performed empirically using the target-decoy approach at both the PSM level and peptide level (using the high scoring PSM as a surrogate for peptide probability). Spectrum level and peptide level q-values were assigned for each PSM. No FDR filtering was performed for MODa results.

Integration of peptide identification results

Data integration was performed by first normalizing the scan numbers output by the different tools. X! Tandem indices were 1 smaller than the ones reported by MS-GF+ and Comet while MODa reports scan numbers as the n-th MS/MS scan present in the MGF file starting at 1. PepXML files and MODa results were parsed and converted to a tab-delimited results file with search results for each spectrum query grouped together and numbered using the MS-GF+/Comet scan numbering.

Screening and interpretation of high scoring decoys

PSMs and peptides were considered a high scoring (PSM and peptide q-values of 0.01 or less) decoy if they matched to a peptide that cannot be found in the forward sequence space for at least one of the three search engines operating in fully tryptic mode. The high scoring decoy is explained as an Ambiguous Scoring event if a high scoring forward hit is identified by some other search engine. If no such tryptic hit is found, the semi-tryptic search results are considered and a Semi-tryptic explanation is assigned if there is a high scoring semi-tryptic forward hit. MODa results are then examined to determine if there is a modified forward peptide that explains the spectrum. If no alternate explanation is found for the high scoring decoy after these steps, it is labeled as Unexplained.

2.2.2 Target-decoy assumption is violated by modified peptides and causes underestimation of error rates

Datasets and data preparation

A deep HEK 293 dataset[65] consisting of 24 LC-MS/MS runs analyzed on a Thermo Scientific Q Exactive mass spectrometer was downloaded from PRIDE (PXD001468). Conversion of vendor .raw files to mzML was performed as previously described.

Peptide identification using MSFragger

Narrow window (100 p.p.m.) and open (500 Da) searching was performed by MSFragger (version 20170103.0) [62] on the HEK 293 dataset against the Ensembl database as described in section followed by peptide validation using PeptideProphet (using the extended mass model) as described in section 3.2.2. Identifications were filtered to 1% FDR at both the PSM and peptide level (using the highest PSM probability).

Generation and database searching of theoretic spectra from modified peptides

The database dump (2014-08-10) of GPMDB[68] was retrieved from their FTP site and parsed using a custom Java program to extract peptide observations from the *peptides* table. The tryptic peptides generated from an in-silico digestion of the Ensembl 78 human protein database was filtered using the list of GPMDB peptide observations to obtain a list of 282,806 peptides. This list of peptides was further filtered to a set of 73,002 peptides by retaining only peptides that have a methionine residue. A random methionine is selected in each of these peptides and is oxidized in-silico and a theoretical spectrum is generated (with carbamidomethylated cysteines) consisting of singly charged b- and y-ions. The theoretical spectra are written to MGF files with the precursor mass and charge reported in the 2+ state.

Peptide identification was performed using the Comet search engine (version 2015.01 rev. 1) against the same Ensembl 78 protein database (both with reversed protein decoys and reversed peptide decoys) using a 20 p.p.m. precursor mass tolerance in high resolution mode (fragment_bin_offset of 0.0 and fragment_bin_tol of 0.02). Fully tryptic digestion was specified with up to 1 missed cleavage and the use of neutral loss ions was disabled for scoring. Variable modifications were disabled as intended and static carbamidomethylation was specified for cysteines. Peptide identifications were ordered by their expectation value and binned into 1000 bins to calculate the fraction of target matches in each bin.

2.3 Results

2.3.1 Characteristics and sources of high scoring decoys

We performed fully tryptic searches using three search engines [34,36,38] and applied a 1% FDR filter at both the peptide and PSM level. The target-decoy strategy assumes that incorrect assignments would match randomly to both targets and decoys. Yet the randomness to which these incorrect assignments are matched to decoy peptides is not well understood and the presence of multiple PSMs supporting the same peptide remains a common filter for additional stringency in proteomics analyses. While the majority of decoy peptides are only supported by a single high scoring PSM, there are many that are supported by tens or even hundreds of PSMs, independent of the search engine used (Figure 2-1). The fact that certain decoy peptides can be supported by hundreds of PSMs suggest that the same is likely to hold true for false positives and that even if a forward peptide is supported by hundreds of high scoring PSMs, it may still be a false positive. The non-uniform nature of these decoy matches across the decoy space indicates that they originate from unaccounted peptides rather than random chemical noise.

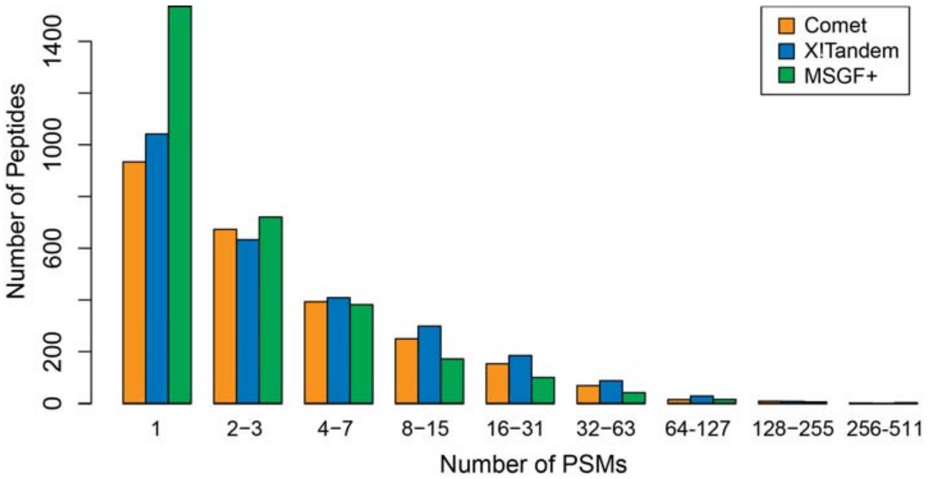


Figure 2-1 High scoring decoys can be supported by tens to hundreds of PSMs.
High scoring decoy peptides are grouped based on their number of supporting PSMs for each of the three search engines operating in fully tryptic mode.

Next, we examined the overlap in high scoring decoy peptide identifications across the three search engines. 7,313 high scoring decoy peptides (at 1% peptide FDR) were identified across the three search engines (Figure 2-2). Surprisingly, only 107 decoy peptides identified by all three search engines. The vast majority (6,576) of high scoring decoy peptides were only identified by a single search engine. The lack of agreement between the different search engines illustrates differences in their spectrum similarity and score calibration functions as the same spectrum is either assigned to another incorrect peptide or fails to reach statistical confidence to pass the 1% FDR filter. This lack of agreement can also be used to improve peptide identifications by combining orthogonal scoring functions in these search engines and removing borderline or conflicting identifications. Indeed, this concept has been used with much success in peptide validation for reducing error rates and improving the number of identifications at a given FDR [69,70].

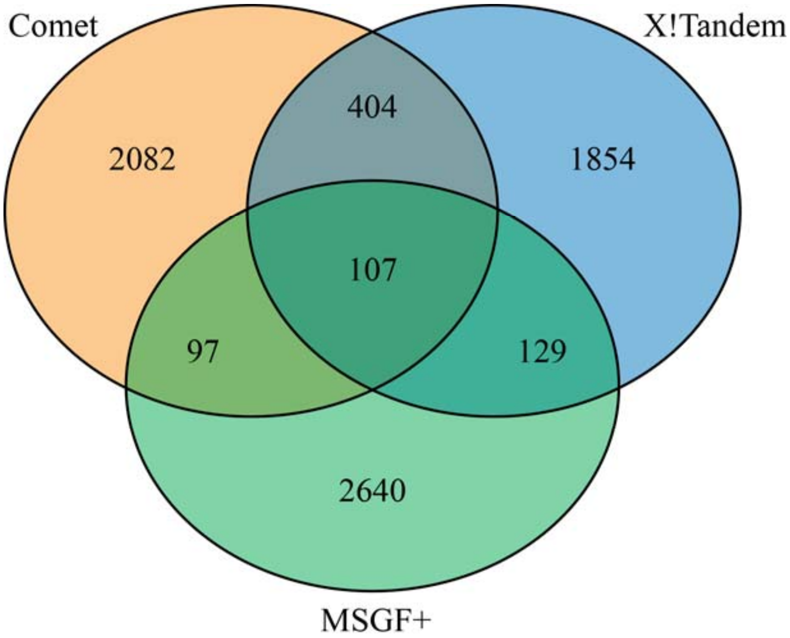


Figure 2-2 Overlap of decoy peptide sequences across three search engines.
High scoring decoy peptide sequences are compared across three search engines. The vast majority of decoy sequences are unique to a particular search engine.

As these incorrect assignments are likely due to unaccounted peptides, we expanded the search space by performing semi-tryptic searches using the three search engines and a blind modification tool [41] to account for unanticipated chemical or biological modifications. We also considered instances where identification is ambiguous – cases where a forward peptide and a decoy peptide are both identified with high confidence by two different search engines. These may represent situations where the database search engine is overly confident in its assignment.

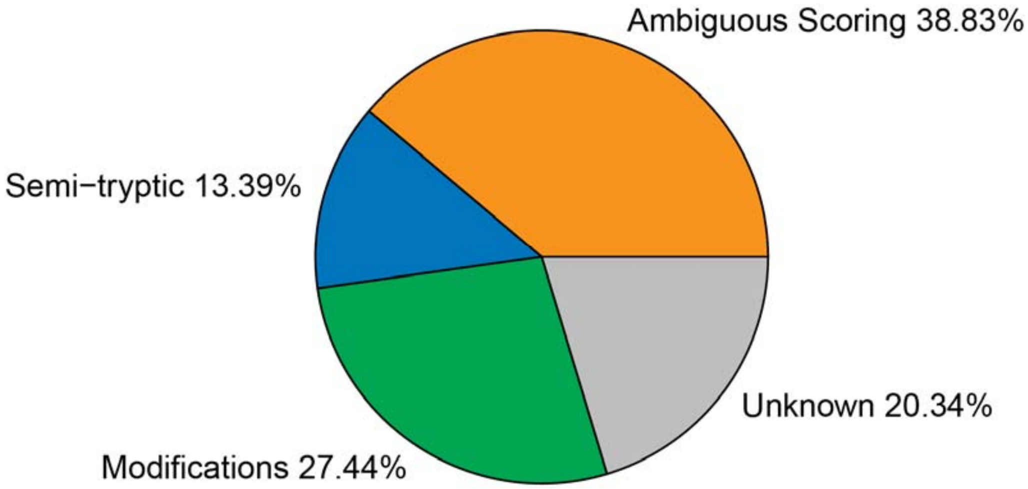


Figure 2-3 Explanation of high scoring decoy PSMs.
High scoring decoy PSMs are iteratively explained using fully tryptic, semi-tryptic, and blind modification search results.

We attempted to find explanations for high scoring decoy PSMs in the following order: ambiguous scoring by the presence of a high scoring forward PSM identified by another search engine, semi-tryptic by the presence of a high scoring semi-tryptic forward PSM identified by any search engine, modified peptide if MODa nominates a modified forward peptide, and unknown if we fail to find a possible explanation for the high scoring decoy PSM (Figure 2-3). In 38.83% of high scoring decoy PSMs, they can be explained by an ambiguous scoring event. For example, the decoy peptide IVESITK was identified with confidence in 286 PSMs. In many cases, the forward peptide LVTDLTK was identified by another search engine. Comparing the matched fragment ions for the two different peptide assignments in one such experimental spectrum (Figure A-1) shows that the matched fragments are identical in mass and that outside of auxiliary information (such as knowledge of the fragmentation pattern or additional fragments from neutral loss ions), the two identifications are of equal confidence and we lack experimental information to distinguish between the two. While the fragments are measured with high mass accuracy and match to a large number of theoretical fragment ions, the presence of multiple peptides that match equally well highlights the difficulty in confidently identifying short peptides.

Next, we examined the population of decoys that can be explained by a semi-tryptic peptide. 13.39% of high scoring decoy PSMs was confidently identified as a semi-tryptic peptide. In our example (Figure A-2), the spectrum matching decoy peptide LANLLVGK was reassigned to the semi-tryptic forward peptide LAGGIIGVK in a semi-tryptic search. The semi-tryptic match improved upon the decoy match by matching two additional y-ions of relatively low intensity. While the correct identification may well be the semi-tryptic assignment, the large number of matched high intensity fragments in the decoy assignment once again demonstrates the limitations of current similarity scoring functions as there is little information to distinguish between the two matches.

We then examined the 27.44% of decoy PSMs that are explained by chemical or biological modifications. We posit that this percentage may be heavily dependent on sample complexity and instrument acquisition speeds. In more complex samples, abundant unmodified peptides dominate the peptide ions that are sampled by the mass spectrometer while less abundant modified forms are ignored. The converse is true for fractionated or low complexity samples with few proteins giving modified peptides a greater chance to be sampled. Increasing the acquisition speed of the instrument or the runtime of the LC-MS/MS run will also likely increase the number of modified peptides sampled and correspondingly, the number of high scoring decoys (and false positives) that are due to unaccounted modifications. In the selected example, the decoy assignment EWHHSHTDITLR fails to match many of the high intensity fragment peaks while the modified peptide assignment of IW[16]HHTFYNELR (oxidation on tryptophan) is of much higher quality, explaining nearly all of the intense peaks in the fragmentation spectrum (Figure A-3). While tryptophan oxidation is well known, it is often not included in routine peptide identification workflows due to the analysis time required to consider additional variable modifications. This demonstrates the utility of blind modification search tools for improving peptide identifications even when the identification of modified peptides is not the primary goal as they have the ability to eliminate false positives. However, many blind modification search tools are slow and incompatible with conventional peptide identification workflows.

Finally, there remains 20.34% of decoy PSMs that cannot be explained by any of the above analysis. As they are assigned to some decoy PSM with high confidence, they contain peptide fragments and are likely of peptide origin. As they cannot be explained as semi-tryptic peptides or modified peptides from the reference proteome, they may be ions with incorrect charge state assignment, peptides derived from alternative splicing [7], peptides with no enzyme specificity [71], or even proteasome spliced peptides [72].

Together, these results suggest that improvements to similarity scoring functions are needed to resolve ambiguous assignments and that proper accounting for all peptide forms (semi-tryptic and modified) is necessary for confident peptide identifications.

2.3.2 Target-decoy assumption is violated by modified peptides and causes underestimation of error rates

High prevalence of peptides identified in modified form only

To further explore the idea that certain false positives are due to modified peptides, we performed both narrow window and open searching (to account for modified peptides) on a HEK 293 dataset (both searches done without variable modifications). We reason that false positives that are due to modified peptides would be confidently identified in the narrow window search (as a false positive) but not in open search (as the supporting spectra would now be assigned to their correct modified peptide identification). To investigate, we looked at the intersection of search results (at the unique peptide level) by subdividing the peptides on the basis of their estimated confidence (Figure 2-4a) and examined the group-specific FDR.

As expected, peptides that were accepted at 1% FDR in both searches (101,138 in total) were of high confidence, with an estimated FDR of 0.15%. Peptides found in both searches but accepted only at 1% FDR in one of the two searches were of lower confidence, as evidenced by the increased group FDR. Of greatest intrigue to us, however, were the peptides that were confidently identified in one search but were not identified at all in the other.

There were 12,622 peptides confidently identified in open search but not in narrow-window search. The relatively low group FDR of these peptides (4.15%) suggests that most of these are bona fide examples of peptides that were only detected in modified forms. The substantial number of such peptides is problematic for ‘dependent-peptide’ approaches for PTM identification [73] (including spectral library-based methods) [74,75] that rely on co-identification of the unmodified peptide. A comparison of the modification profile of these peptides to one that is generated from all modified peptides shows high similarity (Figure 2-4b), suggesting that most of these identifications correspond to constitutive or highly abundant modifications.

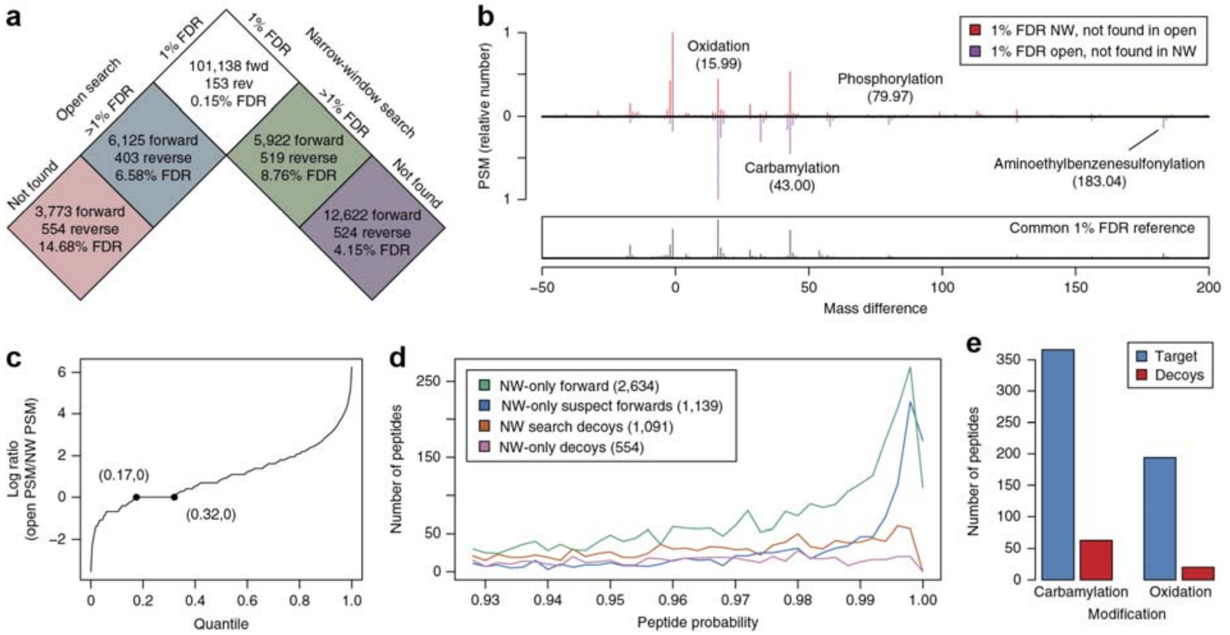


Figure 2-4 HEK293 peptide identifications across traditional narrow-window and open searches demonstrate underestimation of FDR.
Peptides are subdivided on the basis of their estimated confidences in both open and narrow-window search. Group-specific FDR values are estimated using decoys within each group. (b) Mass difference profiles in open search for spectra that identified a peptide unique to narrow-window search (red) or open search (purple) and for all spectra (boxed). (c) Supporting PSM counts in narrow-window and open search for conflicting peptide identifications involving a peptide found only in narrow-window search (at 1% FDR). (d) Comparison of peptide categories passing 1% FDR in narrow-window search. (e) Target and decoy matches in narrow-window search for spectra identified with a common modification in open search.

Open searching uncovers FDR problem in traditional narrow-window searches

In contrast, the 3,773 peptides identified in narrow-window search but not in open search had a much higher group FDR, of 14.68%. We mapped the spectra supporting these identifications to their results in open search. Of particular interest were spectra that were assigned to unmodified peptides in narrow-window search but reassigned, owing to an improved match, as modified peptides (with different sequence) in open search. These cases represent potential instances of false positives in narrow-window search that are caused by chemical or biological modifications [76,77]. In each such instance—a pair of peptides whose masses differ by the mass of the modification detected in the open search—we compared the total number of supporting PSMs associated with the peptide sequence matched in narrow-window search to that in open search (Figure 2-4c). Assuming that peptides supported by a greater number of PSMs are more likely to be true identifications, we found substantially more support for the peptides identified in open search. Only 17% of the spectra were assigned to peptides that had greater support in narrow-window search, whereas 68% had greater support for their open-search assignment.

We called peptide identifications found only in narrow-window search to be ‘suspect’ (potential false positives) if there was greater support for the open-search assignment for each supporting PSM. Of the 3,773 peptides found only in narrow-window search, 1,139 were suspect. This is significantly more than the number of decoys (554) in the same group, and more than the total number of decoys in the entire narrow-window search, at 1% FDR (1,091 decoys in total). This suggests that false positives in narrow-window search are not correctly estimated by decoy peptides. Notably, some of these suspect peptides had very high scores (Figure 2-4d).

We sought to verify the finding that the target–decoy strategy does not effectively capture false positives due to unaccounted modifications. We selected high-scoring peptide identifications in open search that were observed in both unmodified form and with a mass shift corresponding to a common modification (oxidation or carbamylation). As we did not specify any variable modifications, the target-decoy assumption is that spectra from these modified peptides would match equally (and incorrectly) to both targets and decoys in narrow-window search. However, that was not the case, as the rate of matching to target sequences was roughly six fold that of decoys for carbamylated peptide spectra, and more than nine fold for oxidized peptide spectra (Figure 2-4e). The violation of the target–decoy assumption is probably due to homology between true peptide sequences and other peptides in the target space, which we previously noted in the context of proteogenomics [7,76]. Further supporting this, the modification profile of peptides identified in open search and whose spectra produced suspect identifications in narrow-window search markedly lacked phosphorylation and aminoethylbenzenesulfonylation (Figure 2-4b). These two mass shifts (79.97 and 183.04 Da) are difficult to represent, as some sequence of amino acid addition and deletion. Overall, our analysis with the HEK293 data set demonstrates that accounting for all modified peptide forms using the open-search strategy of MSFragger may be important for confident peptide identification, even when the identification of modified peptides is not the primary interest.

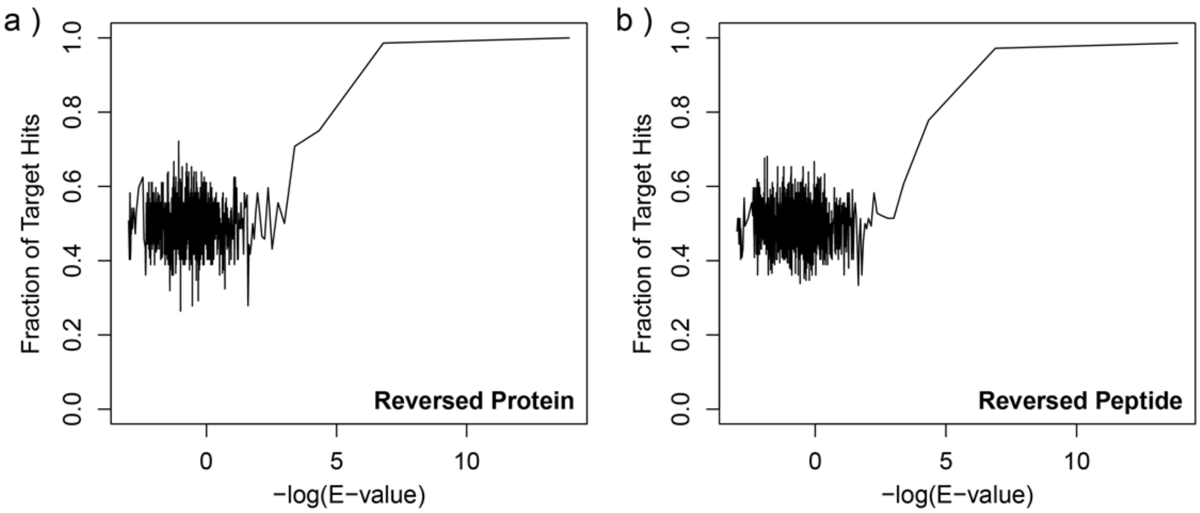


Figure 2-5 Spectral homology in theoretical spectra derived from modified peptides.
PSMs of theoretical spectra derived from a peptide with oxidized methionine are ordered by their expectation value and binned to calculate the fraction of target hits as a function of expectation value (a) Analysis performed using reversed protein sequence as decoys (b) Analysis performed using reversed peptide sequences as decoys

Validation of target-decoy violation using theoretical modified spectra

While the searching of experimental spectra identified as modified peptides in a narrow window search suggested that the target-decoy assumption is violated, we must proceed with caution due to the confounding problem of chimeric spectra. In some cases, a modified peptide (the top identification in open search) can be co-fragmented with an unmodified peptide which is then identified (correctly) in the narrow window search that does not account for modifications. To circumvent the problem of chimeric spectra, we generated theoretical spectra for a set of peptides with oxidized methionine. As expected, for the majority of these theoretically pure spectra, they map to both targets and decoys at equal rates. However, for the highest scoring matches, there is a strong preference for target sequences, regardless of the method used in decoy generation (Figure 2-5). Unsurprisingly, some of the highest scoring PSMs are due to single amino acid differences that are equal to the mass of the added oxygen atom (e.g. alanine to serine in M[16]ASTFIGNSTAIQELFK matches MSSTFIGNSTAIQELFK with e-value 3.62E-20). There are also examples where the peptides differ in the number of amino acids but remain spectrally similar (e.g. M[16]DVNVGDIDIEGPEGK matches NSHHSWEPLDAPEGK with e-value 4.71E-8). Together, these results show that modified peptides can be potential sources of false positives that are not well modeled by the target-decoy strategy.

2.4 Discussion

In this chapter, we demonstrate that nearly 80% of high scoring false positives can be explained through additional computational analysis. Nearly half of the explainable false positives can be attributed to ambiguous identifications where there is a lack in search engine concordance. This suggest that combining results from multiple search engines [69] or similarity scoring functions might help in identifying these cases and reducing their assigned confidences. It also points to limitations in the scoring functions presently used in database search engines where only a rudimentary model for predicting theoretical spectra is used (equal ion intensities). The use of spectral libraries generated from synthetic peptides [78] or fragment ion intensity prediction programs [32,79] may help resolve such ambiguities and reduce the rate of such false positives in the future. The high confidences with which these ambiguous identifications are reported may also be a consequence of the availability of high mass accuracy data in both MS1 and MS2, where a few matched fragment ions result in high confidences due to a reduction in fragment matches of a purely random nature. The other half of explainable false positives are due to peptides that are excluded (semi-tryptic and modified peptides) from searches due to computational costs or otherwise complexity in integrating different search results. While the computational costs associated with database searching can be readily addressed, new statistical models may need to be implemented to integrate multiple search spaces, each with different prior probabilities and selecting the explanation with highest posterior probability, to avoid a loss in sensitivity when vastly expanding the search space.

The comparison between open and narrow window search results provides several insights into the properties of modified peptides and their impact on the production of false positives and error rate estimation. Over 10% of the peptides identified were identified only in modified forms, suggesting that methods which can directly identify modified peptides without co-identification of the unmodified peptide (such as tag-based PTM search tools or the open search strategy) are likely to be the most successful in comprehensive identifying modified peptides and reducing the number of false positives. Second, both the experimental and theoretical results indicate, to varying degrees, that the target-decoy assumption is violated for modified peptides and that modified peptides are more likely to match to target sequences than decoy sequences, causing an underestimation of false positives that are due to modified peptides. The true extent of this violation is of much interest and further experiments using experimental spectra while accounting for the effects of co-fragmentation are needed. It is also important to note the limitations of this study and its generalizability to proteogenomics results. Open modification searching identifies only 50% of common modifications when compared to direct interrogation using specified variable modifications [65]. The underreporting of modified peptides may imply that the number of false positives due to modified peptides may be greater than what was shown in this study. Furthermore, the database searched in this comparison was the Uniprot protein database with few variant peptides or peptides that are highly homologous to one another compared to a database produced in a sequencing experiment. Hence, the error rates for such variant peptides could potentially be higher than what is estimated here. Experiments with decoys designed to contain single amino acid substitutions (confirmed not to exist from sequencing data) would allow us to directly interrogate errors of this nature.

2.5 Data availability

Raw mass spectrometry files are available from public repositories as described. The processed data files supporting the findings of this study are available upon request.

2.6 Acknowledgments

This work was performed by Andy Kong, under the supervision of Dr. Alexey Nesvizhskii.

CHAPTER III

EFFICENT DATABASE SEARCH TOOLS FOR PROTEOGENOMICS ANALYSIS

Portions of this chapter detailing the MSFragger algorithm have been published in Nature Methods [62]

3.1 Introduction

Database search has long been a bottleneck in computational proteomics workflows [20] with continual efforts to improve the speed of identifications to match the pace of instrument acquisition speeds and the growth of protein databases. In proteogenomics experiments, the database to be searched is nearly 10X larger than a reference database, increasing the search time accordingly. Due to the already large search space that decreases sensitivity and long search times, semi-tryptic searches and multiple variable modifications are not commonly used. As demonstrated in the previous chapter, this can lead to an accumulation of false positives when peptide species present in the sample are unaccounted for in the search space. Further compounding this problem is the increasing popularity of the open search approach for identifying blind modifications. While simple in its approach (a matter of changing the precursor tolerance to hundreds of Daltons), it is a brute force approach, comparing each experimental spectrum to hundreds of thousands of candidate peptides. This leads to analysis times of over 50 central processing unit (CPU) hours (per LC-MS/MS run) [62] using conventional database search tools, making it costly for large scale analyses. Hence, there is need for faster, more efficient database search tools for proteogenomics analysis.

Computational advances in database search tools often fall in two orthogonal avenues. The first utilizes more advanced or sophisticated computing hardware such as parallel computing technologies such as networked compute clusters [80–82] to distribute the computation workload on conventional processors or repurposing specialized hardware such as graphical processing units (GPUs) found in commodity graphics cards [83–85]. The second involves algorithm improvements that reduce the amount of computation needed to obtain peptide identifications of similar quality. This can involve strategies for using heuristics to filter the search space [36,40], improved implementations of scoring functions [86,87], or peptide indexing methods that eliminates redundant peptides and recycles the results of in-silico digestion [88,89]. In this chapter, we present two database search tools. The first, EGADS, utilizes GPUs to accelerate both in-silico digestion and similarity scoring while the second, MSFragger, uses a novel fragment ion-indexing scheme to vastly improve the speed of spectra similarity calculations.

Algorithms that utilizes networked compute clusters can perform high-throughput searches but do not reduce the overall amount of computation time required considering all CPUs, making them costly and financially prohibitive for large-scale analyses. GPUs have been used as a more efficient alternative in many scientific applications as they offer the compute capabilities of hundreds of CPUs in a graphics card costing several hundred dollars. These GPUs consist of thousands of processing cores coupled to fast onboard memory. However, specialized algorithms are needed to function on GPUs due to the limited memories and single instruction multiple data (SIMD) nature of GPU cores where each individual processing cores must perform the same operation in lockstep but on potentially different pieces of data. Due to these complexities, GPUs have only been applied to the spectra similarity scoring step of database search [83,85] which has is more easily parallelizable and is historically the more computationally intensive step of database searching with low precursor mass accuracies. However, with high resolution instruments that records precursor masses with accuracies on the order of several parts per million [26,27], in-silico digestion may become a bottleneck when only similarity scoring is accelerated. Hence, EGADS implements both GPU accelerated digestion and scoring providing significant speedups regardless of the precursor mass accuracy.

The growing popularity and ease-of-use of the open searching strategy for identifying modified peptides precipitated the development of MSFragger. Open searching using a 500Da window can take nearly 1000X longer than traditional narrow searches with a precursor tolerance in the tens of p.p.m. [62]. The use of traditional search engines is computationally prohibitive on large datasets and GPUs are not widely or inexpensively available from cloud providers. Hence, using the relatively large amount of system memory, we developed a fragment ion indexing scheme that allows similarity scoring to be rapidly performed simply by traversing the index using experimental fragment ions.

3.2 Materials and methods

3.2.1 EGADS an Efficient GPU-Accelerated Database Search tool

EGADS development environment

EGADS was developed in C++ using Visual Studio 2010 on Windows 7 64-bit. OpenCL support was provided by the AMD APP SDK (version 2.9). Testing was performed on a desktop computer equipped with an Intel 2500K processor with 16GB of memory with an AMD Radeon 7950 GPU with 3GB of memory (using the latest AMD drivers as of May 2013).

Datasets and data preparation

A HeLa dataset [22] consisting of 3 technical replicates each with 6 fractions analyzed on a Thermo Scientific Orbitrap Elite was downloaded from PRIDE (PXD002395). Vendor raw files were converted to MGF using ProteoWizard as previously described. The human protein database was obtained from Refseq (release 55) and reversed protein sequences were appended as decoys. For the scenarios involving a hypothetical proteogenomics use case, an mRNA database was obtained from Refseq (release 55) and was translated into a protein database using three-frame translation. ORFs shorter than 10 amino acids were discarded in this translation.

EGADS algorithm

(1) Spectra input and pre-processing

EGADS reads MS/MS spectra in MGF format and pre-processes them according to the chosen similarity scoring function after filtering the input spectrum to the specified top N peaks after filtering out peaks with m/z greater than 2048 (as EGADS calculates the Xcorr function using 2048 bins). For the Xcorr scoring function, an offset of 0.4 Da is added to the m/z of each input peak before binning to unit m/z bins (the maximal intensity is taken when there are multiple peaks that fall within the same bin). The 2048 bins are then subdivided into 16 equal windows and intensities within each window are normalized so that the maximum intensity within each window is 50. The intensities are then transformed into the form necessary for fast cross-correlation [86] by computing cumulative sums. For PeakMatch and PeakBackground, no binning is performed. PeakMatch normalizes the peak intensities so that the maximum intensity is 100, while PeakBackground performs the local normalization as performed in Xcorr using the ranges of 64 m/z but without binning.

(2) Trial digestion and scoring

EGADS performs in-silico digestion and simulated scoring (where the number of similarity scores required is calculated but the scoring is not actually performed) on a small region of the database (20 blocks of 4096 characters) to estimate the amount of GPU memory necessary to perform on-GPU digestion and scoring for a given amount of sequence space. EGADS then uses this estimate and the available GPU memory (after accounting for the space necessary to store the database and experimental spectra) to partition the sequence space that can be independently analyzed in multiple digestion-scoring cycles (Figure 3-1).

(3) In-silico digestion and peptide de-duplication

EGADS performs in-silico digestion by first concatenating all proteins (separated by end-of-protein delimiters) into a single string representing the entire sequence space. The string is then subdivided into blocks of 4096 characters that can be analyzed by a single workgroup. Individual threads scan overlapping ranges of 128 characters in order to determine the number of digested peptides that starts within the first half of its range. These counts are then summarized in scan operations and used to allocate appropriate memory for storing masses and offsets of these digested peptides. The process is repeated to store the digested peptide products.

Peptides are then optionally de-duplicated (within the same digestion-scoring cycle) to eliminate redundant peptides and the number of repeated similarity scoring calculations. We re-pack the digested sequence into a 160-bit integer that is unique for all peptides less than 32 amino acids in length (the upper limit in EGADS). This integer is then iteratively sorted 32 bits at a time using a GPU-efficient version of radix sort [90] (which is also used for other sorting operations within EGADS) grouping redundant peptides together in this sorted list.

Next, peptides are expanded into modified mass peptides that represent the different masses that arise as a result of user specified variable modifications and are sorted by mass. For each modified mass peptide, we also compute the number of structural isomers. Peptides that produce more than 100 (hard coded limit that can be changed) modified versions of that peptide are considered not modified (modifications are ignored). Together, these elements form a peptide index for the current digestion-scoring cycle.

(4) Similarity scoring and result reporting

In GPU accelerated scoring, all threads within the same workgroup operate on the same experimental spectrum. The number of structural isomer, along with the number of experimental spectra that have a mass within the precursor mass window, are used to compute the number of workgroups necessary for scoring. This information is then used to launch one of three scoring kernels that are responsible for computing the peptide-spectrum similarity scores. Lookup tables are used to efficiently map permutations to modified residues.

Three scoring kernels were implemented that mimics modes in popular database search tools: Xcorr (unit-resolution Comet [34]), PeakMatch (computes hyperscore as in X! Tandem [36]), and PeakBackground (high-resolution Comet). Xcorr is computed a simple product after spectra pre-processing described above while PeakMatch and PeakBackground is calculated using sliding windows to merge the peak lists between the experimental and theoretical spectrum in time linear to the total number of peaks.

Similarity scores are transferred back to the CPU where a histogram of scores and a heap of the top hits are maintained for each experimental spectrum. The similarity scores are converted in a scoring function dependent manner to an expectation value. Results are written to a pepXML file that is compatible with PeptideProphet by masquerading as either Comet or X! Tandem.

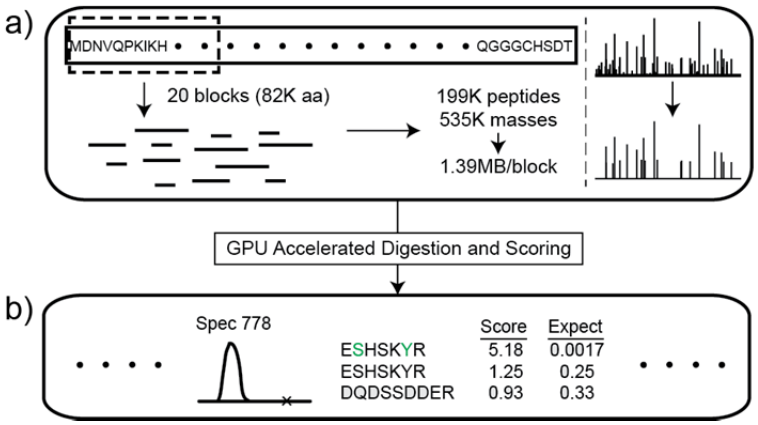


Figure 3-1 Architecture of EGADS.
EGADS performs trial digestion and spectra pre-processing on the CPU before starting digestion-scoring cycles on the GPU. (b) Results are transferred back to the CPU where the histogram of scores for each spectrum is modeled and an expectation is calculated for the reported hits.

Timing and Performance Measurements

Timers are placed throughout EGADS to collect timing information on the cumulative time spent in individual steps of the database search process. Database search implementations can vary widely so comparing runtimes between different database search engines is not meaningful for determining the extent of GPU acceleration. Hence, EGADS implements identical data structures and algorithms used in the OpenCL kernels in C++ (referred to as EGADS CPU) so that the same parallel algorithms that are run on the GPU can be run serially on the CPU. EGADS does not make use of CPU-GPU concurrency as all OpenCL calls made are blocking.

For timing purposes, we consider digestion to be the steps in the digestion-scoring cycle up to the point of counting structural isomers and scoring to be remaining steps in the cycle, including update of the results structure. Input, initialization (including compilation of the OpenCL kernels if no cached version exists), spectra conditioning and pre-processing, and output takes no longer than several seconds in total and is not a major factor in overall runtime.

Benchmarking of EGADS on a HeLa dataset

Peptide identification benchmarking was performed by searching the Refseq 55 protein database (described above) allowing for 1 missed cleavage in fully tryptic digestion. Static carboxyamidomethylation (+57.021464 Da) on cysteine was specified while allowing oxidation (+15.9949 Da) on methionine to be a variable modification. The permitted peptide mass range was 500-2000 Da (results were filtered to exclude matches outside this range in X!Tandem where the maximum cannot be specified). The precursor mass tolerance was set to 100ppm with fragment mass tolerance set to 20ppm (0.02 Da bins for Comet). Isotope error correction was disabled. The use of neutral loss ions was disabled in Comet. For EGADS and X! Tandem, the top 50 fragment ions was used in searching. EGADS was given access to 2GB of memory using the PeakMatch scoring kernel. Run time benchmarking was performed using the above settings (except Comet was run in unit-resolution mode for the comparison with Xcorr), varying the digestion mode, precursor window, and scoring kernel. Apart from the identification rate benchmarking, all analyses was performed on the run 20100611_Velos1_TaGe_SA_Hela_1.

Peptide validation was performed using PeptideProphet [48] (TPP 4.8.0) using the options d' (report decoy hits), 'A' (high mass accuracy model), 'P' (semi-parametric modeling), 'PPM' (parts per million in mass model) and FDR was estimated empirically using the target-decoy approach after ordering PSMs and peptides (represented by highest scoring PSM) by PeptideProphet probabilities.

Combinatorial evaluation of GPU acceleration

Evaluation of GPU acceleration in diverse search applications was performed by varying the following search parameters: de-duplication (on / off), phosphorylation search (on / off), precursor mass tolerance (20 ppm / 1 Da), and sequence space (Refseq 55 protein / 3-frame translated Refseq 55 RNA), digestion mode (tryptic with 1 missed cleavage, semi-tryptic with 1 missed cleavage, non-specific), and similarity function (Xcorr / PeakMatch / PeakBackground). All other search parameters were set to the values described above for benchmarking. EGADS was run in both CPU and GPU mode for each of the 144 search combinations. Runs taking longer than 4 hours on the CPU was killed and excluded from the analysis (66 out of the 144 combinations completed successfully).

Benchmarking memory effects on EGADS runtime

For evaluation of the memory effects on EGADS runtime, the Refseq protein database described above was searched using EGADS running in GPU mode using all three digestion modes (fully tryptic, semi-tryptic, and non-specific). For each memory limit (1024MB, 1536MB, 2048MB, and 2560MB), five runs were performed and the average runtime calculated after discarding the shortest and longest runs.

Open database searching using EGADS

The PeakMatch scoring algorithm was modified to include two additional ion series: b + Δ and y + Δ to account for any shifted fragment ions due to a modified residue in open searching. Δ is computed as the difference between the experimentally observed mass and the theoretical peptide mass.

MODa (version 1.23) was used in single blind mode with fragment mass tolerance set to 0.02 and high-resolution MS/MS. In both EGADS and MODa, static carboxyamidomethylation was specified. Variable oxidation of methionine was specified in EGADS but all results involving oxidized methionine were filtered out in both searches as there it was not possible to specify variable modifications in MODa. The allowed modification range was -200 - +200 Da in both searches. No FDR filtering was performed in the open search analysis.

3.2.2 MSFragger implements a novel fragment ion index that enables ultrafast database search

Datasets and Data Preparation

A HEK 293 dataset[65] consisting of 24 LC-MS/MS runs analyzed on a Thermo Scientific Q Exactive mass spectrometer was downloaded from PRIDE (PXD001468). Conversion of vendor .raw files to mzML was performed as previously described.

MSFragger algorithm

(1) MSFragger spectra input and pre-processing

MSFragger accesses mzXML and mzML files using MSFTBX, the Data Access Library provided as part of the BatMass [91] and Mascot Generic File (MGF) files using an internal parser. These data input paths allow MS/MS spectra stored in any of the three file formats (mzXML/mzML/MGF) to be analyzed by MSFragger. Spectra pre-processing begins with linear scaling of peak intensities so that the most intense peak within each spectrum is set to 100,000. Resultant scaled intensities are rounded and stored as integers for fast arithmetic operations. The top N peaks from each spectrum are retained and are then filtered based on the minimum intensity ratio and the m/z range specified in the search parameters file. In this study, the top 100 peaks with a minimum intensity ratio of 0.01 (relative to the base peak) were used with no m/z range filter.

(2) In-silico protein digestion and peptide indexing in MSFragger

MSFragger allows for fully enzymatic, semi-enzymatic, and non-enzymatic digestion to be specified as search parameters. It also allows for limits on missed cleavages, peptide lengths and masses to be specified. For a given protein database and a fixed set of digestion parameters, a peptide index is generated to form a necessary reference for the fragment index. Peptide indexing takes just a few minutes on a typical computer. Furthermore, MSFragger caches the peptide indices it generates on disk and attempts to find and use a compatible peptide index on subsequent invocations. As the first step of in-silico digestion, all proteins are concatenated into one long amino acid sequence with proteins separated by delimiter characters. MSFragger then partitions this long amino acid sequence into chunks for parallel in-silico digestion into peptide sequences based on the specified digestion parameters. Efficient memory allocation methods and compact representations of peptides (as offsets in the concatenated amino acid sequence and length) allow for fast in-silico digestion. The digested peptide sequences are then sorted using a parallel least significant digit radix sort and redundant peptides are flagged by comparing adjacent peptide sequences in the sorted list of peptides.

Modified versions of the digested peptide sequences are then generated based on the user-specified variable modifications. Combinatorial bitmasks that specify the positions of modified residues are pre-computed so that the set of variably modified residues can be specified as a single integer. These sequence numbers are then combinatorially combined across all variable modifications to generate a single integer that represents the variable modification state of a peptide sequence. A 12-byte entry containing the offset, length, modification sequence number, and the modified mass is generated for each such modified peptide. These modified peptides are then sorted in parallel by their modified mass forming the MSFragger peptide index.

(3) Fragment index generation

The fragment index used in MSFragger consists of all theoretical b and y-ions up to a specified charge state from each peptide in the peptide index. For efficient fragment index searching, the fragment bin width used for the fragment index must be proportional to the desired fragment tolerance specified in the search and to the expected number of candidate peptides encountered per experimental spectrum. Hence, MSFragger dynamically computes an appropriate bin width, in Daltons, that allows for efficient fragment index searching based on the user specified precursor mass tolerance and the fragment mass tolerance. Each peptide entry in the peptide index, consisting of both unmodified and variably modified peptides, can be referenced by a single 32-bit integer identification number (ID), imposing a current limit of approximately 2 billion peptide entries. Within each peptide entry, the theoretical fragments are generated and binned based on their masses using the determined bin width. The theoretical fragments are stored within the fragment index as an 8-byte entry that references the parent peptide ID, the mass offset within the bin, the charge state, and the fragment ion identity (e.g. b-5 or y-2). Fragments within each bin are stored in order of their parent IDs (and hence the parent precursor mass) as the fragment index is generated in the order of the peptide index. The memory consumption of the fragment index is modest. For a tryptic digestion (with 1 missed cleavage) of the human UniprotKB database (with reversed decoys) used in the study, the fragment index is only 1.6GB. Adding methionine oxidation and N-terminal acetylation of proteins as variable modifications boosted the index size to 2.9GB. Examples of fragment index sizes (which includes the above common variable modifications) for larger search spaces include HLA peptides (non-enzymatic digest of 9-11 amino acids; 22.6GB), semi-tryptic peptides (55.8GB) and variably phosphorylated peptides (86.5GB). MSFragger identifies the amount of memory available to it via the Java Virtual Machine and automatically partitions the fragment index generation and search into multiple iterations based on projected memory required for the fragment index, storing intermediate results on disk before merging and outputting the results in the final pass. This enables MSFragger to perform searches on computers that do not have sufficient memory to store the full fragment index, although at reduced speeds. In addition to the fragment index, MSFragger requires additional memory for storing the peptide index, spectra data, results, and intermediate data structures during search that is roughly 1GB in most use cases.

(4) Fragment index searching

In database search, the similarity scores are computed between each experimental spectrum and the theoretical spectra of all candidate peptides within a precursor mass range. These scores are heavily dependent on the number of shared fragment ions between the experimental spectrum and theoretical spectra. The major computational advance presented by MSFragger lies in its ability to rapidly identify these shared fragment ions and thus compute spectrum-spectra scores with near optimal efficiency. MSFragger first identifies the number of candidate peptides using the precursor mass window and the computed peptide index. It then allocates a scoring table for each candidate peptide where the number and summed intensities of matched b and y-ions can be stored. It then performs spectrum to spectra scoring using the fragment index in the following manner. Consider a fragment ion with mass M within an experimental spectrum with precursor mass P. Using the fragment index, the algorithm can identify the theoretical spectra that contain a fragment with a matching mass by examining the fragment bins that overlap the interval [M – dF, M + dF], where dF is the fragment mass tolerance specified in Daltons or otherwise computed from M and the specified tolerance in parts per million (Figure 3-10).

For each overlapping fragment bin, a binary search (recall that the fragments within each bin are ordered by their parent precursor masses) is used to identify the fragment within the bin that corresponds to precursor mass P – dP, where dP is the precursor mass tolerance. The bin is then traversed and the theoretical fragments within the bins are compared to determine whether they truly lie within the fragment mass tolerance window, and if the theoretical fragment charge state is compatible. If a match is identified, the scores of the parent peptide (recall that each theoretical fragment contains a reference to its parent) are then incremented in the scoring table. This traversal continues until the end of the bin or upon arrival at a fragment with parent precursor mass greater than P + dP. The process is then repeated for each overlapping fragment bin. At completion, this process using a single experimental fragment ion represents the contribution of that fragment ion to all spectrum-spectra scores. This process is repeated for each experimental fragment ion (Figure 3-10), in essence, decomposing many spectrum-spectrum matches into multiple fragment-spectra matches. After processing all experimental fragment ions, the scoring table of candidate peptides contains the number of matching ions (and intensities) and is used to generate a similarity score for each candidate peptide.

The efficiency of this process lies in its ability to only examine fragments with a high likelihood of contributing to the similarity score. In conventional strategies, performing a comparison between an experimental spectrum and a theoretical spectrum can take tens or hundreds of operations, even in cases where they share no common fragments. In the MSFragger strategy, theoretical spectra that share no common fragments are effectively bypassed (apart from reading a score of 0 from the scoring table) as mostly relevant fragments are compared. In the case of open window searching, approximately 1.5 comparisons are performed on average per candidate peptide and over 80% of fragment comparisons within the fragment index contribute to a similarity score (Figure 3-2). This algorithmic advantage that allows MSFragger to perform so few comparisons in similarity calculations is the reason why it performs over 100 times faster than conventional search tools.

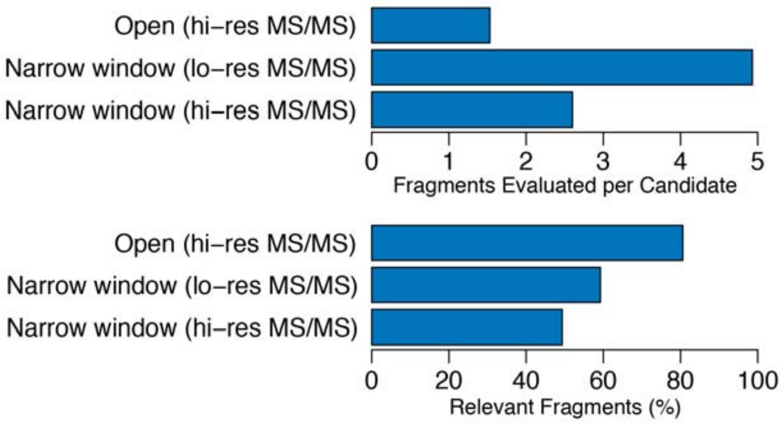


Figure 3-2 Fragment indexing allows efficient spectra similarity comparisons.
The cost and efficiency of spectra similarity calculations can be approximated by the number of fragment comparisons required for each candidate peptide. In conventional strategies, tens to hundreds of comparisons are needed to compare an experimental spectrum to a theoretical spectrum. However, the vast majority of such fragment-fragment comparisons do not result in matches as the differences between their m/z is often far greater than the fragment mass tolerance. Using MSFragger's fragment index, these comparisons are omitted as the binning strategy allows us to retrieve only the experimental-theoretical fragment pairs that are close in m/z – the majority of which falls within the fragment mass tolerance and are deemed relevant when they contribute to the score of a PSM. MSFragger's alternative approach results in only a few fragments evaluated per candidate peptide across a variety of search scenarios. Reduction in the fragment bin width allows for fewer fragment comparisons to be performed at the expense of computational overheads associated with traversing a greater number of bins that overlap the fragment tolerance window. MSFragger dynamically selects a bin width appropriate for the search scenario (opting for smaller bins in open search where the number of comparisons is large, and larger bins in narrow window search, where the number of comparisons is small relative to the overhead costs). Hence, a greater number of fragments is evaluated per candidate and a lower percentage of comparisons are found relevant in narrow window searching due to this optimization.

Fragment index searching in MSFragger is highly optimized. Tradeoffs between the number of bins to traverse (cost of binary searching and other overhead) and hit efficiency (percentage of fragments that fall within the fragment mass tolerance) is weighted and considered in fragment bin width selection. The traversal algorithm is optimized for modern CPU cache sizes to reduce main memory accesses using a simultaneous traversal scheme for all experimental fragment ions. This allows for overall improved performance and reduces memory bottlenecks in multi-core systems (Figure 3-3).

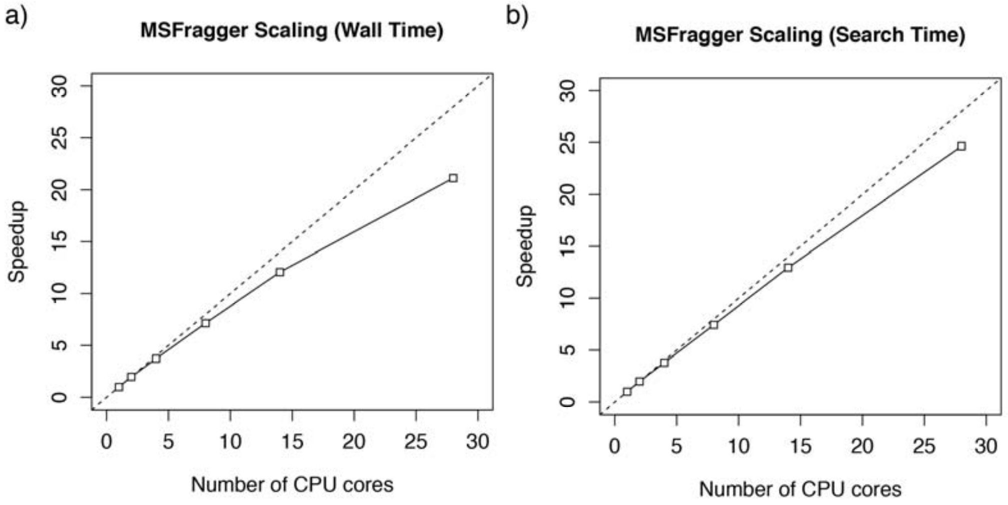


Figure 3-3 MSFragger scales efficiently across large numbers of CPU cores.
Indexing and searching operations in MSFragger are designed for modern multi-core computers and are optimized to reduce pressures on memory bandwidth. Results are generated from open search times of a single LC-MS/MS run on a dual processor system with 14-cores in each processor. (a) MSFragger scales almost linearly in terms of overall search times on up to 8 cores. Reading of mass spectrometry data files and results compilation is not highly parallelizable resulting in reduced scalability beyond 8 cores. The jump from 14 to 28 threads causes non-local memory to be accessed by each processor, impacting scalability. (b) Fragment index searching by itself is efficiently parallelizable in MSFragger and scales to effectively utilize all cores.

(5) Scoring and results reporting

MSFragger computes a hyperscore similar to that of X!:

$$hyperscore = \log(N_b!\,N_y! \sum_{i=1}^{N_b} I_{b,i} \sum_{i=1}^{N_y} I_{y,i})$$

where N_b is the number of matched b-ions, N_y is the number of matched y-ions, $I_{b,\copyright}$ are the intensities of matched b-ions, and $I_{y,\copyright}$ are the intensities of matched b-ions. While the theoretical fragment index can be adapted to include other fragment ion types, only b and y ions are included and used for scoring at this time. Expectation calculation is also performed in a similar manner as X! Tandem through linear regression of the survival function to estimate the expectation of a given hyperscore [92]. The top N results, as specified by the search parameters, are reported in a XML file in the pepXML format, which can then be processed using the tools from the Trans-Proteomics Pipeline (TPP) [67]. For use in other computational workflows, converters exist that can convert pepXML results into other standard data output formats. Alternatively, a simple tab separated values output of the results can be obtained instead of the pepXML.

Statistical modeling of MS/MS search results and protein inference

X! Tandem, Comet, and MSFragger output files were uniformly processed by PeptideProphet [48] via the Trans-Proteomic Pipeline (TPP v4.8.0), followed by ProteinProphet [49] analysis to assemble peptides into proteins/protein groups. The results from the narrow window searches were processed using the following settings: PeptideProphet was run using ‘P’ (semi-parametric modeling), ‘d’ (report decoy hits), ‘E’ (calculation of posterior probabilities using search engine computed expectation values as primary peptide identification scores), and ‘A’ (high mass accuracy model), ‘PPM’ (use parts per million instead of Daltons in accurate mass binning), and the ProteinProphet was run using default settings. For open searches, several custom modifications were made to these downstream processing tools. First, PeptideProphet was run without ‘A’ and ‘PPM’ options, and using a mass accuracy model extended to cover the entire (-1000Da to 1000Da) range (see Extended mass model below). Second, in ProteinProphet, we did not want to incorporate modified peptides in the determination of protein groups or the establishment of protein identities. Thus, ProteinProphet was adjusted to ignore any modified peptides, while being careful to retain peptide identifications that are likely triggered from C13 isotope peaks of unmodified peptides.

Extended mass model in PeptideProphet

For open searches, the mass model of PeptideProphet was extended to effectively adjust for different likelihoods of obtaining a correct identification among unmodified peptides and peptides with different types of modifications (mass shifts). In brief, PeptideProphet models the distribution of scores observed in each data set as a mixture of two component distributions representing correct and incorrect identification, respectively. The key underlying assumption is a multivariate mixture distribution of the database search score (here, the expectation values produced by the search tools) and other parameters (most notably, the mass shift *dM*), which leads to the calculation of the probability of correct identification for individual peptide assignments by the Bayes rule. The mass shift parameter *dM* (which in the context of narrow window searches is referred to as mass accuracy) is computed for each PSM as the difference between the calculated and measured precursor peptide masses [93]. Unlike narrow window searches, in open searches the range of possible *dM* values is extended, e.g. to cover (-1000 Da to 1000 Da) range. The *dM* values are discretized into bins of 1 Da in size (centered at integer values). The distributions of database search scores and *dM* mass shifts are modeled simultaneously, resulting in the joint probability model and computation of posterior peptide probabilities. In doing so, the mass shift *dM* model is estimated from the data, defining likelihoods of observing a correct vs. incorrect identification among all PSMs belonging to a particular *dM* bin. As the main outcome, two PSMs with identical expectation values but having different binned dM values (e.g. 0 and 135) would receive very different probability scores, reflecting the fact that the estimated fraction of correct identifications in the *dM* ~ 0 bin (i.e. unmodified peptides) is much higher than that among peptides with a *dM* value around 135 Da (rare modification). Note that while the mass model helps to account for the differences in the likelihoods of observing unmodified peptides and different modified forms, coarse single Dalton binning fails to account for the parts per million (ppm) levels of accuracy present in these data from high mass accuracy instruments, and thus the model can further benefit from future revisions.

Benchmarking analysis using HEK293 dataset

For extensive benchmarking and comparison between MSFragger and other tools using HEK293 dataset, all spectra were searched using MSFragger, X! Tandem (Piledriver 2015.04.01.1), and Comet (2015.02 rev.1). Analysis was done using all files (24 LC-MS/MS runs, ~1.1 million spectra) for identification rate benchmarking, or one representative file for timing benchmarks (run b1906, 41820 spectra). The searched sequence database was created from the human protein sequences of Ensembl version 78 appended with reversed protein sequences as decoys and common contaminants (cRAP proteins sequences from gpmDB and contaminants from MaxQuant). All searches were done considering only *y*- and *b*- ions in scoring, allowing tryptic peptides only, up to 1 missed cleavage, and with cysteine carbamidomethylation specified as a fixed modification. Data were searched using either 100 ppm (narrow windows searches) or 500 Da (open searches) precursor mass tolerances. X! Tandem search engine used the following algorithm-specific parameters: select top 50 peaks for fragment matching, 20 ppm fragment ion mass tolerance, and requiring at least 4 matched fragment ions for a PSM to be reported. Note that X! Tandem automatically considers three additional modifications (conversion to pyroglutamate from glutamine or glutamic acid, and N-terminal acetylation). Comet searches were performed using recommended settings for high mass accuracy fragment data (precursor mass binning of 0.02 Da, 0 mass offset). MSFragger searches were performed using described parameters. To enable more accurate comparison with X! Tandem results, MSFragger searches (both narrow window and open) were also performed allowing same common modifications as those mentioned above for X! Tandem specified as variable modifications. For comparison with SEQUEST, the identification numbers (as listed in Table 3-3), i.e. the numbers of PSMs, unique peptide sequences, and proteins, were taken from the original publication [65].

For benchmarking the computational time (as listed in Table 3-3), MSFragger, Comet, and X! Tandem were also run using the single representative file referenced above on a quad core Linux workstation (Intel Xeon E3-1230v2). In addition, the data were searched using Tide (Crux version 2.1.16838), which only allows a maximum of 100Da mass tolerance and is single threaded. The run time for Tide, and for MSFragger run under the same constrains as Tide, are shown in Table B-1. For SEQUEST, the computational time listed in Table 1 was obtained by searching the data using the SEQUEST-HT version as implemented as part of the Proteome Discoverer v. 2.1 software, operated on a octa-core workstation (2x Intel Xeon E5-2609v2). The search parameters for SEQUEST-HT were as above, except the mass tolerance in the narrow window search was 5 ppm as in the original publication. All computational time benchmarking results can be found in Table B-1.

Comparison between MSFragger and MODa

MODa (v. 1.51) was run in single-blind mode with a maximum modification size of 500 Daltons and a fragment tolerance of 0.02 Daltons. Cysteine carbamidomethylation was specified as a static modification. High resolution MS/MS search was enabled. Tryptic digestion was specified with at most one missed cleavage. Both fully tryptic and semi-tryptic searches were performed using MODa. FDR filtering was performed using the “anal_moda.jar” tool bundled with the MODa tool to achieve a FDR of 1%. For comparison with MSFragger, we filtered the fully tryptic MSFragger open search results at 1% PSM FDR (without the 1% protein level filter that was used for the rest of the HEK293 benchmark comparison).

3.3 Results

3.3.1 EGADS an Efficient GPU-Accelerated Database Search tool

Framework for GPU-accelerated in-silico digestion

In traditional database search tools, peptides are generated one at a time by traversing the protein database and experimental spectra within the precursor mass window are identified and scored. This process is simple and has minimal memory requirements but lacks the parallelism for efficient acceleration. In contrast, peptide indexing methods [37,94,95] decouples the in-silico digestion from similarity scoring by creating a sorted index of unique peptides prior to similarity scoring. These methods have the advantage of eliminating redundant peptides and allow results of in-silico digestion to be re-used in subsequent searches if the index is stored (which may not be practical or efficient due to their size). In EGADS, we developed a hybrid strategy that takes advantage of the massive parallelism offered by the GPU to generate on-the-fly peptide indices that can be directly used for similarity scoring on the GPU in multiple digestion-scoring cycles (Figure 3-4).

EGADS begins by partitioning the sequence space into chunks that can be processed within the limited GPU memory. It then performs in-silico digestion of the given chunk into individual peptides given the specified cleavage rules. These peptides then undergo an optional de-duplication step to remove redundant peptides and are then expanded into multiple entries representing each modified masses that can be generated from variable modifications. The entries are then sorted to create the peptide index. Candidate peptides are then identified for each experimental spectrum and the number of scoring events is computed (by summing the number of structural isomers for each modified mass). The similarity scores are then computed in parallel and results are transferred to the host. At this point, the peptide index, along with all intermediate data structures, is discarded to free up GPU memory for the next digestion-scoring cycle (Figure 3-5). At the end of all digestion-scoring cycles, the results are summarized on the host and written to a file.

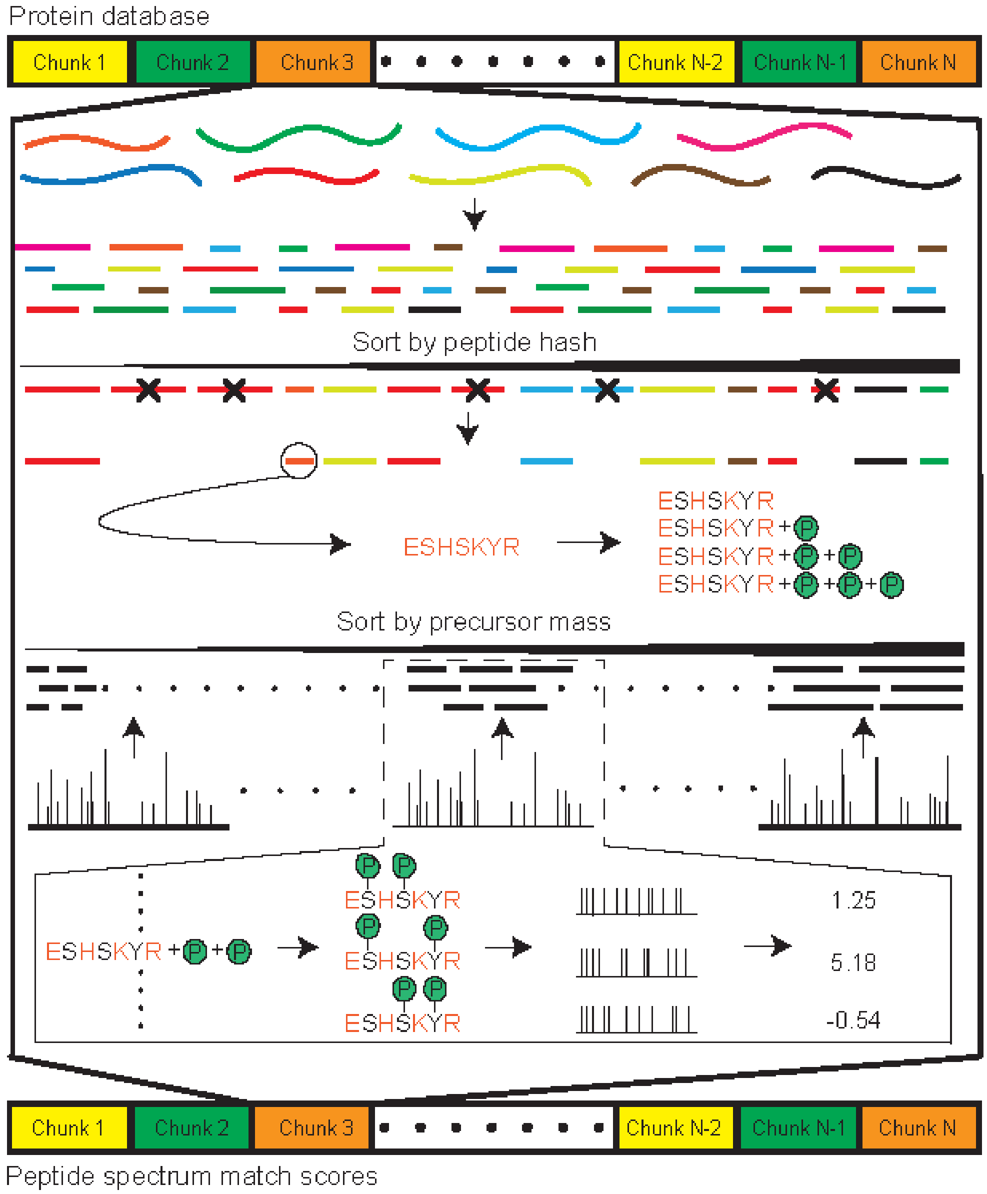


Figure 3-4 Digestion-scoring cycles in EGADS.
In each digestion-scoring cycle, the partitioned database chunk is transformed in parallel to a peptide index on the GPU. The peptide index is used for similarity scoring on the GPU. After the scores are reported to the host, the index is discarded to begin the next digestion-scoring cycle.

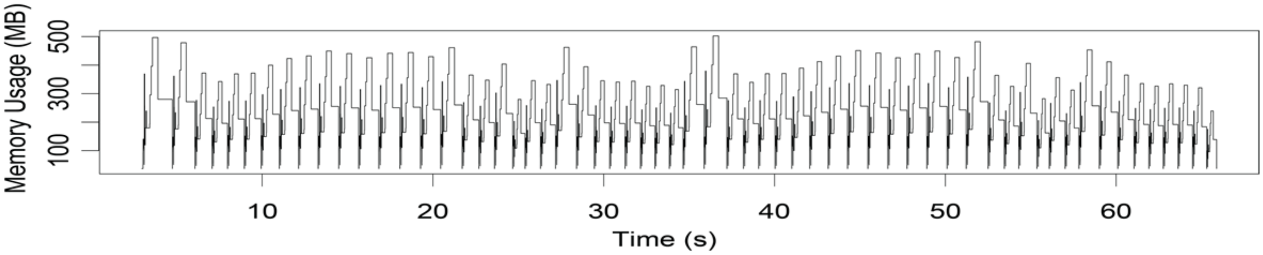


Figure 3-5 GPU memory usage of EGADS in digestion-scoring cycles.
Memory allocation by EGADS is shown for a typical EGADS run. Each peak and trough represents one digestion-scoring cycle.

Benchmarking of EGADS using a HeLa dataset

We first compared the number of identifications that is obtained by EGADS to that of Comet and X! Tandem using a HeLa dataset. At 1% FDR, EGADS (using the PeakMatch kernel) identified 45,248 peptides (unique peptide sequences) and 225,228 PSMs. Comet and X! Tandem respectively identified 46,620 and 49,866 peptides, and 238,978 and 212,408 PSMs (Figure 3-6). Paradoxically, X! Tandem identified the fewest number of PSMs but the highest number of peptides. This might be explained in part by the fact that X! Tandem has certain variable modifications that cannot be disabled which allows it to identify a greater number of peptides but the sensitivity of the cross-correlation function allows Comet to identify the greatest number of PSMs (with EGADS performing similarly as X! Tandem for PSMs as it also used the PeakMatch function but performs similar to Comet at the peptide level as it does not account for these additional variable modifications). Overall, EGADS performs similarly to these popular search engines and can reliably perform peptide identifications.

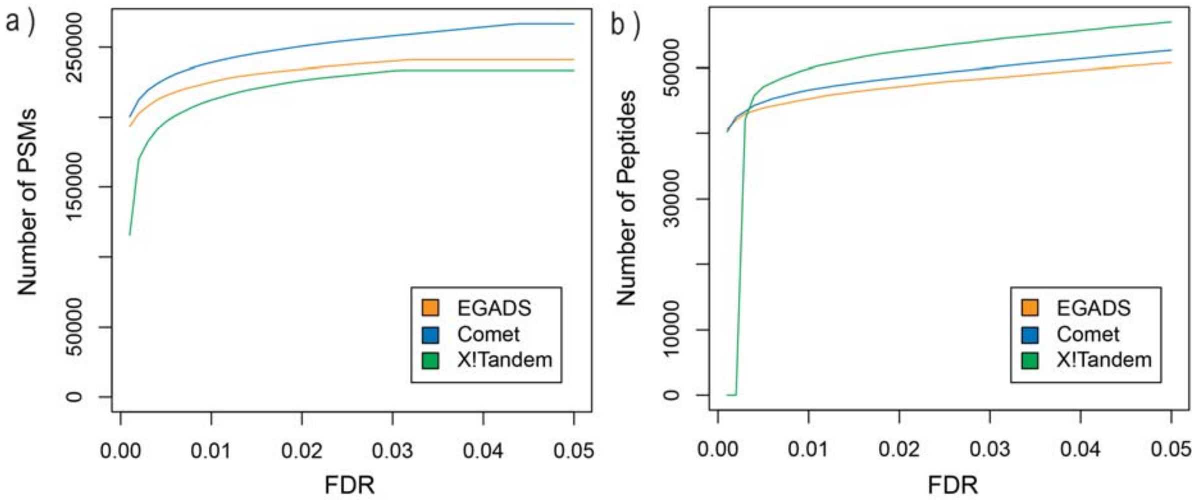


Figure 3-6 Identifications in the HeLa dataset as a function of FDR.
Number of peptide-spectrum matches as a function of FDR. (b) Number of identified peptides as a function of FDR.

We then evaluated the runtime performance of EGADS, which is the primary focus of this study. We compared EGADS to Comet and X! Tandem using the Xcorr and PeakMatch functions respectively under different digestion modes and precursor tolerances (Table 3-1). Comet only takes 1.6 times longer to perform a semi-tryptic search when compared to a fully tryptic search. Given the large difference in search space, this suggests that Comet spends a substantial amount of time in operations unrelated to digestion and scoring (such as input pre-processing). In contrast, EGADS takes 16.4 times longer to perform the semi-tryptic search suggesting that it is much more efficient in input pre-processing. On average, EGADS GPU is 36.3X faster than Comet but much of that can be attributed to differences in input pre-processing as EGADS GPU is only 4.3X faster than EGADS CPU. This highlights the importance in having a reference CPU implementation to ensure that implementation differences are not attributed to GPU-acceleration. In the PeakMatch comparison, EGADS CPU performs similarly as X! Tandem, with EGADS being more efficient in input pre-processing and X! Tandem being slightly more efficient in digestion and scoring (based on the relative changes in runtimes as the search space increases). On average, EGADS GPU performed 36.9X faster than X! Tandem and 25.7X faster than EGADS CPU, showing that PeakMatch is much more amenable to GPU-acceleration than the Xcorr similarity function.

Table 3-1 Running time of EGADS compared to Comet and X! Tandem.
Overall running time in seconds of EGADS in both CPU and GPU mode compared to popular search tools.

Search engine	Search time (seconds)			
	Tryptic 20ppm	Tryptic 1Da	Semi-tryptic 20ppm	Semi-tryptic 1Da
		Xcorr performance		
Comet	363.0	381.0	582.0	2379.0
EGADS CPU	11.2	37.2	184.0	889.2
EGADS GPU	4.9	10.0	30.0	172.6
		PeakMatch performance		
X! Tandem	161.3	512.3	1104.1	7606.6
EGADS CPU	41.7	302.8	1023.7	8517.2
EGADS GPU	4.3	11.2	34.6	233.1

Effects of search parameters on GPU-acceleration

To further investigate the influence of various search parameters on GPU-acceleration, we ran EGADS in both CPU and GPU mode across a diverse set of conditions, recording the total time taken in individual steps across all digestion-scoring cycles (Table B-2). We first looked at the acceleration of the digestion step across the three digestion modes (Figure 3-7). Average speedup for tryptic digestion was 15.1X, semi-tryptic digestion was 13.4X, and non-specific digestion resulted in an average speedup of 13.1X. The trend of reduced performance gains for more complex digestions is likely due to the overheads associated with each digestion-scoring cycle. As the complexity of the digestion increases, the database must be partitioned into smaller chunks so that the intermediate data structures can fit within the same amount of GPU memory. This phenomenon is also affected by the precursor mass window. Average digestion speedup of 15.0X was observed for runs using a precursor mass window of 20ppm while an average of 13.1X was observed for runs with a precursor mass window of 1 Da. The increased precursor mass window translates into a larger number of scoring events per digested peptide. The memory required to calculate these additional similarity scores result in fewer peptides processed per digestion-scoring cycle, leading to larger number of cycles and reduced acceleration.

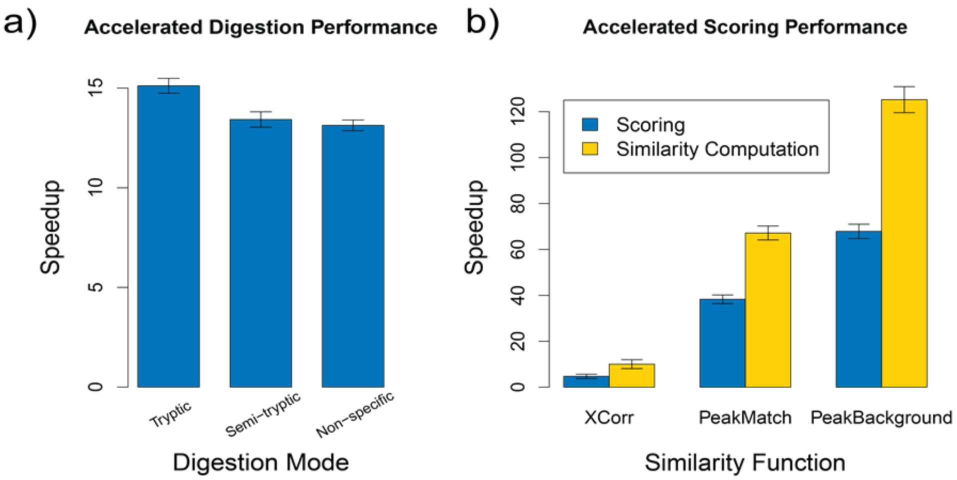


Figure 3-7 Average GPU speedups as a function of digestion mode or similarity function.
Speedups are computed relative to the reference EGADS CPU implementation. (a) Average speedup of all digestion steps with standard error bars shown. (b) Average speedup for the entire scoring phase and for the similarity computation step only. Standard error bars are shown.

We next examined the acceleration of the similarity scoring. The scoring phase is composed of three main steps: preparation of scoring blocks, computation of similarity scores, and updating the per-spectrum results structure. For the entire scoring phase, we observed average speedups of 4.7X for Xcorr, 38.3X for PeakMatch, and 67.8X for PeakBackground. Focusing only on similarity score computations resulted in speedups of 10.0X for Xcorr, 67.2X for PeakMatch, and 125.2X for PeakBackground (Fig 3-7). These trends reflect the differences in scoring function complexity with greater speedups for more complex functions where the work to data ratio is large. For example, the Xcorr function is calculated as a simple dot product between the theoretical spectrum and the pre-processed experimental spectrum. The amount of information (the binned experimental spectrum) transferred to the local processor is large compared to number of operations performed (sums up the bins where theoretical fragments exist) and quickly saturates the memory bandwidth on the GPU, limiting gains. In contrast, the PeakMatch and PeakBackground algorithms use peak lists, which are much more compact compared to a binned spectrum, and performs many more operations (checking fragment tolerances as the two peak lists are merged) in scoring, leading to much large speedups compared to the reference CPU implementation. As a result, even though there are great differences in the computational complexity across the three scoring functions, their actual runtimes are quite similar after GPU acceleration (Figure 3-8).

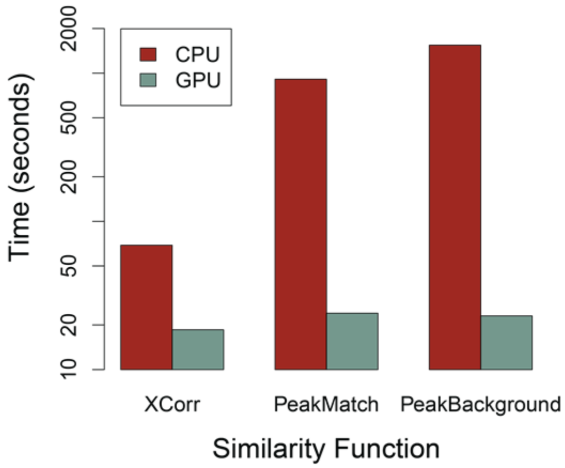


Figure 3-8 Total time used for similarity scoring for different similarity functions.
Total amount of time spent in the similarity scoring phase for searches using a semi-tryptic digestion of the Refseq protein database and a 20 ppm precursor mass tolerance.

In CPU based scoring, the computation of similarity scores accounts for over 90-99% of the computation time, depending on complexity of the scoring function (lower for simple functions like Xcorr). However, in GPU accelerated scoring only the computation of similarity scores is accelerated leading to situations where it takes longer to update the results structure than to perform similarity scoring on the GPU (notably when Xcorr is used). Introduction of CPU-GPU concurrency where the CPU performs this task while the GPU works on the next digestion-scoring cycle can eliminate this bottleneck in future work. We also note that the time taken for CPU-based digestion is 2.76 times that of GPU-based scoring, indicating that digestion, rather than scoring would be the bottleneck if not for the GPU-accelerated digestion implemented in EGADS. This is particularly important for high-resolution data as this ratio increases to 3.41 if instances with a 1 Da precursor window are excluded.

Impact of GPU memory on EGADS runtime

As memory plays a large role in effecting the speedup of GPU accelerated searches, we directly examined its effect on the overall runtime of a GPU accelerated search at different memory limits. This is also important for informing GPU purchasing decisions as the same GPU can be sold with differing amounts of on-board memory. We performed tryptic, semi-tryptic, and non-enzymatic searches of the Refseq protein database using a precursor tolerance of 20ppm (the effects at 1 Da are more pronounced as more memory is required for similarity scoring but is an unlikely use case for modern high resolution instruments) and the PeakMatch algorithm (Table 3-2). For fully tryptic searches, EGADS does not benefit from additional memory as it can accommodate the entire database in 1-2 digestion-scoring cycles. For semi-tryptic search and non-enzymatic search, moving from 1024MB to 2560MB decreased the runtimes by 15.7% and 11.1% respectively. This modest improvement indicates that while EGADS benefits from additional GPU memory, its scheme for partitioning the computation allows for effective GPU acceleration even on devices with limited on-board memory.

Table 3-2 Running time of EGADS as a function of on-board memory.
EGADS is operated in GPU mode using the PeakMatch algorithm in each of the three digestion modes. Total search times are recorded for different amounts of allocated GPU memory.

Digestion Mode	Search time (seconds)			
	1024MB	1536MB	2048MB	2560MB
Tryptic	7.93	7.89	7.89	7.89
Semi-tryptic	9.76	9.01	8.51	8.23
Non-enzymatic	200.11	190.21	182.37	177.87

Application of EGADS to open database searching

We investigated the application of EGADS for blind modification searching by adapting the PeakMatch algorithm to include additional ion-series that accounts for shifted fragment ions and running it in open mode (precursor mass tolerance set to 200 Da). For comparison, we performed blind modification searching using MODa [41], an established blind modification search tool that uses sequence tags. On a single run from the HeLa dataset, MODa took 7333 seconds to perform a single-blind modification search. In contrast, EGADS took 357 seconds using its modified PeakMatch scoring kernel, making it 20.5 times faster than MODa in nominating the most abundant modifications. The recovered modification profiles were similar, identifying common expected mass differences (Figure 3-9). Interestingly, MODa appears to have a slight bias towards smaller mass differences based on the modification profile of decoy sequences (which can be assumed to be uniform in nature).

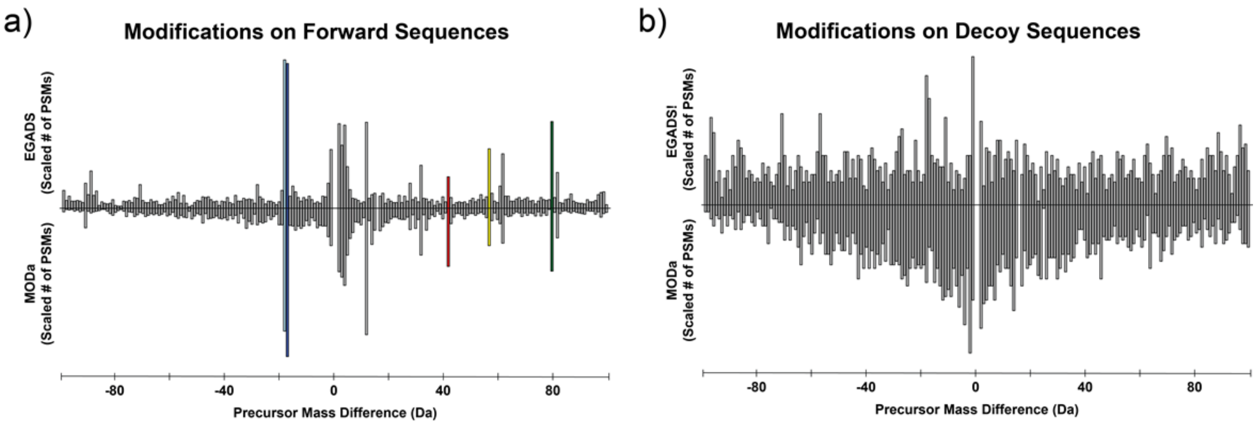


Figure 3-9 Modification profiles of HeLa dataset as determined by EGADS open searching and MODa.
Scaled histograms of PSM counts binned to unit mass differences. Bins 0 (unmodified), 1 (missed assignment of mono-isotopic peak), and 16 (oxidation) are hidden. (a) Modification profile comparing EGADS and MODa on forward sequences. (b) Modification profile comparing EGADS and MODa on decoy sequences.

3.3.2 MSFragger implements a novel fragment ion index that enables ultrafast database search

Novel fragment ion index enables ultrafast database search

MSFragger begins by performing an in silico digestion of the protein database (Figure 3-10). It then removes redundant peptides and orders them by their theoretical mass (including any modified peptides generated as a result of variable modifications), creating a peptide index. Although peptide indexing has been described as a way to accelerate database search [37,94,95], this step alone has little impact on spectrum similarity calculations, which is the most time-consuming step. MSFragger addresses this bottleneck by creating a novel theoretical fragment index. This enables highly efficient and simultaneous scoring of an experimental spectrum against all candidate peptides (see section 2.2.2, Figure 3-2 and Figure 3-10 b-d).

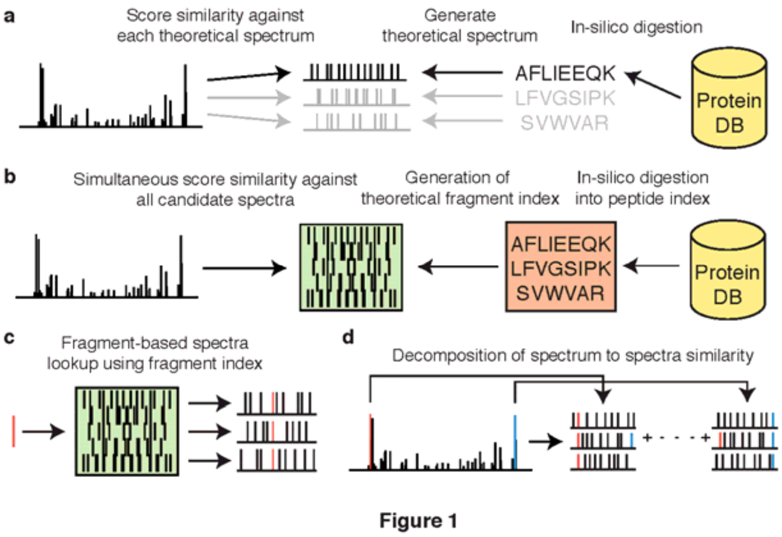

a
Score similarity against each theoretical spectrum
Generate theoretical spectrum
In-silico digestion
AFLIEEQK
LFVGSIPK
SVWVAR
Protein DB
b
Simultaneous score similarity against all candidate spectra
Generation of theoretical fragment index
In-silico digestion into peptide index
AFLIEEQK
LFVGSIPK
SVWVAR
Protein DB
c
Fragment-based spectra lookup using fragment index
d
Decomposition of spectrum to spectra similarity
Figure 1

Figure 3-10 Database-search strategies and the MSFragger algorithm.
Conventional database search involves in silico digestion of a protein database (DB) into candidate peptides from which theoretical spectra are sequentially generated and compared against experimental spectra. (b) MSFragger digests a protein database and generates a nonredundant set of peptides that are arranged in an index, which is then used to generate a fragment index for efficient and simultaneous scoring of an experimental spectrum against all candidate spectra. (c) Mass binning and precursor mass ordering in the fragment index allows rapid retrieval of candidate spectra that match a given experimental fragment ion. Scores of candidate peptides corresponding to retrieved spectra are incremented. (d) Processing of experimental fragment ions results in identification of all matching fragments between the experimental spectrum and candidate theoretical spectra, decomposing spectrum-to-spectra matches to fragment-to-spectra matches. Matched fragments can then be used to compute a similarity score.

We first evaluated the performance of the MSFragger algorithm on a deep HEK293 [65] data set and compared it to that of commonly used search engines Comet [34] and X! Tandem [36]. The scores and error rates of modified peptides are likely to be different from those of unmodified peptides, prompting class-specific FDR estimation [96,97]. To account for these differences, we adopted an extended mass model when computing peptide probabilities, ensuring mass-shift-dependent FDR estimation and filtering. We note that in open searches, the term ‘modifications’ is used interchangeably with ‘mass shifts’ and includes in-source fragmentation events, missed cleavages, and isotope errors. Overall, all search engines performed similarly when run using similar search parameters (Table 3-3). In the traditional (narrow-window) search, MSFragger and Comet identified 9,795 and 9,757 protein groups (1% protein FDR) and 456,548 and 461,806 PSMs (1% protein and PSM FDR), respectively. MSFragger also identified similar numbers as X! Tandem, when accounting for the innate variable modifications that X! Tandem specifies by default. In open search, which represents the primary motivation for the development of MSFragger, we observed a dramatic increase in the number of identified PSMs across all search engines, in line with the earlier report using SEQUEST4. For example, MSFragger identified 609,897 PSMs using open search, an increase of 33.6% compared to narrow-window search, with a minimal loss of 1.4% in the number of protein identifications. When performing protein inference using open-search results, we took a conservative approach, using only unmodified peptides and peptides with specified variable modifications (Online Methods). When all modified peptides were included, the number of protein identifications from open searches exceeded that of narrow-window searches (for example, by 4.4% for MSFragger). However, additional work is necessary to carefully evaluate the accuracy of the protein inference step when using all peptides identified in open search.

Table 3-3 Identification rates and analysis time for the HEK 293 data set.
Identification numbers are for the entire 24-run LC-MS/MS data set, filtered at 1% FDR at both protein and PSM levels. Search times given are for a single LC-MS/MS run consisting of 41,820 MS/MS spectra analyzed on a quad-core workstation.

Search engine	Time (min)	Proteins	PSMs	Peptides
		Narrow-window search		
SEQUEST[a]	9.3	9,513	396,736	110,262
Comet	1.7	9,757	461,806	115,612
X! Tandem[b]	1.7	10,182	466,701	119,304
MSFragger	0.4	9,795	456,548	115,755
		Open search (500 Da)		
SEQUEST[a]	673.0	9,178	510,139	111,205
Comet	815.4	9,545	584,218	123,679
X! Tandem	976.0	9,830	638,052	133,318
MSFragger	5.4	9,656	609,897	126,037

[a] For time estimation, SEQUEST searches were performed using Proteome Discoverer 2.1 (SEQUEST HT) on a more powerful eight-core workstation. Narrow-window searches were done with a 100-p.p.m. precursor mass window, except for SEQUEST (5 p.p.m.). SEQUEST identification rates were taken from ref. [65].
[b] X! Tandem searches include several variable modifications that cannot be turned off.

Open searches using conventional database search tools are slow, given the vastly expanded search space. Comet and X! Tandem took 13.6 and 16.3 h, respectively, to analyze a single LC-MS/MS run using a quad-core workstation. In stark contrast, MSFragger took only 5.4 min, making it >150 times faster than these commonly used tools. We also compared MSFragger to tools that employ peptide indexing, such as Tide [89] and SEQUEST HT (Table B-1). Tide, which allows 100-Da precursor windows only and does not take advantage of multiple processor cores, took 176.7 min (compared to 9.8 min with MSFragger when subjected to the same constraints). SEQUEST HT (Proteome Discoverer 2.1) took more than 11 h on a more powerful octa-core workstation. The speed and scalability (Figure 3-3) of MSFragger allowed open searching of the entire HEK293 data set (24 LC-MS/MS runs) in less than 30 min on a single powerful workstation, compared to the days or even weeks that would be required to search these data using existing tools on the same machine.

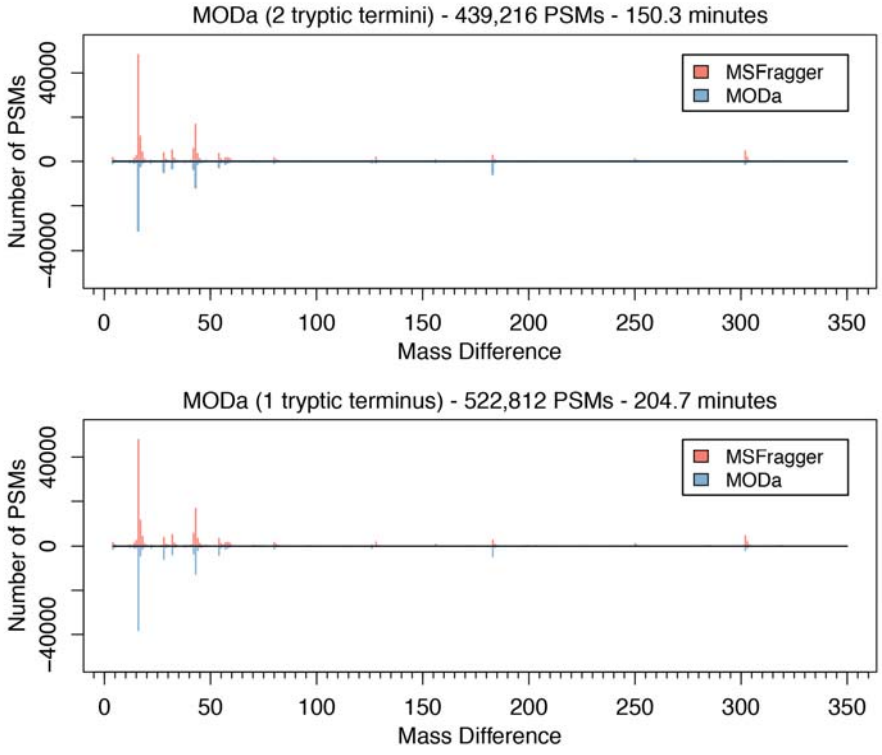


Figure 3-11 Open searching identifies similar modifications as MODa.
MODa, run in single blind mode, generates a similar modification profile as that of an open search with differences that are likely due to the characteristics of the modification. Open searches (run in fully tryptic mode in both comparisons) are more likely to recover mass shifted peptides that have little discernible alterations in their tandem mass spectra (such as the modification near 302 Da) as it does not attempt to localize the modified mass. MODa is likely more effective for modifications that are more commonly found near the C-terminus (and disrupts the y-ions used in open search identification). MODa running in semi-tryptic mode (the mode of operation as recommended by its authors) recovers a greater number of PSMs at the expense of additional run time.

We also sought to compare MSFragger to algorithms specifically designed for comprehensive PTM analysis. MODa[41] has been established as an effective tool for blind PTM search. Using comparable settings, both tools produced very similar PTM profiles (Figure 3-11), but MSFragger identified a larger number of PSMs than MODa, at an FDR of 1%: 622,857 (MSFragger, tryptic search) versus 522,812 (MODa, semi-tryptic search) and 439,216 (MODa, tryptic search). The difference between MODa and MSFragger results can be explained, in part, by the fact that MODa's algorithm localizes the mass shift to a particular amino acid, whereas open searching identifies only the peptide sequence and the mass shift (which may be the result of multiple modifications). Considering the computational time, MSFragger was notably faster (5.4 min; 24.5 min when restricted to a single core) than MODa, which took 204.7 min (semi-tryptic search) and 150 min (tryptic search) for a single LC-MS/MS run on a quad-core workstation (Table B-1).

3.4 Discussion

In this chapter, we presented two database search tools that are both considerably faster than any existing tools. EGADS allows narrow window searches to be rapidly performed on large search spaces (such as semi-tryptic or non-enzymatic digestions involving custom databases predicted from sequencing data) as the GPU-accelerated digestion component is much faster than the CPU-based indexing performed by MSFragger while not performing a large number of similarity scores. However, the requirement of specialized hardware makes it inaccessible to many users and prohibits deployment on inexpensive cloud computing resources. The ability of MSFragger to run on conventional computers makes it much more accessible and scalable. Its algorithmic strength in similarity calculations makes it ideal for open searches where a large number of similarity comparisons are calculated. However, for narrow window searches, much of the fragment index can be left untouched when they represent the theoretical fragments of digested peptides that have no experimental spectra within the precursor mass tolerance. This inefficiency can be ameliorated to a certain extent by running MSFragger in batch mode, where multiple LC-MS/MS runs are processed sequentially, reusing the generated fragment index and increasing the coverage of the m/z range by experimental precursors. For users without access to GPUs, MSFragger in narrow window mode still enables much faster searches than conventional search tools, especially in batch mode. However, for labs equipped with the capability to perform both types of searches, the two tools or approaches are complementary, with EGADS more suitable for narrow window searches and MSFragger for open searches. It is also worthy to note that EGADS implements multiple scoring functions, potentially enabling multiple GPU accelerated searches to be performed using different scoring functions to reduce the number of false positives that are ambiguous identifications.

The indexing algorithm presented in MSFragger has the potential to be used in a number of different mass spectrometry applications where comparing spectra is a computational bottleneck. This includes spectra clustering, testing new hypotheses on indexed experimental spectra, or identifying spliced peptides. The algorithm can also be adapted to consider shifted fragment ions to further empower the open search concept by building different fragment indices based on fragment charge. Finally, as the approaches are orthogonal, a GPU implementation of MSFragger is theoretically possible, allowing such an implementation to be thousands of times faster than the conventional database search tools today that do not employ either innovation

3.5 Data availability

Raw mass spectrometry files are available from public repositories as described. The processed data files supporting the findings of this study are available upon request. The EGADS database search tool is available upon request. MSFragger can be obtained from www.nesvilab.org/software. A software manual for running MSFragger can be found in Appendix D.

3.6 Acknowledgments

The EGADS study was performed by Andy Kong, under the supervision of Dr. Alexey Nesvizhskii. The MSFragger project was conceived by Andy Kong and Dr. Alexey Nesvizhskii. Andy Kong developed the algorithm, wrote the software and analyzed the results. Dr. Alexey Nesvizhskii assisted with algorithm development, software design, analysis of the results, and supervised the entire project. Felipe da Veiga Leprevost, Dmitry Avtonomov, and Dattatreya Mellacheruvu contributed to software development and data analysis.

CHAPTER IV

COMPREHENSIVE PROFILING OF MODIFIED PEPTIDES IN SHOTGUN PROTEOMICS USING THE OPEN SEARCH STRATEGY

Contents of this chapter have been published in Nature Methods [62]

4.1 Introduction

Post-translational modifications (PTMs) regulate cellular functions in ways that cannot be studied through deep genomic or transcriptomics sequencing. A wide range of PTMs including phosphorylation, ubiquitination, glycosylation, and acetylation has been characterized and localized to tens of thousands sites in the human proteome [98]. The most common and prevalent PTM studied is phosphorylation likely due to their function in signaling and involvement in cancer processes [99]. In these studies, the PTM of interest is often chemically enriched prior to analysis by mass spectrometry. Hence, it was of surprise and excitement when it was reported that an additional 30% PSMs were identified to be modified peptides by the open search strategy in an non-enriched HEK 293 proteome, indicating that there is a vast assortment of chemical and post-translational modifications that can be studied without enrichment in datasets already present in public repositories [100] representing diverse tissues and conditions. Furthermore, the open search strategy is known to recover only 50% of modified peptides, suggesting a potential increase in the identification rates by another 30%. The large number of modified peptides collected in non-enriched samples might be due to recent increases in instrument speed, allowing less abundant modified peptide forms to be effectively sampled in data-dependent acquisition. These modified peptides have long eluded identification in tandem mass spectrometry even though their spectra are collected due to the computational burdens of identifying blind modifications, making them part of the dark matter of shotgun proteomics [101]. The open search strategy provides a simple and direct opportunity to illuminate this dark matter and allow us to comprehensively study the full complement of chemical and post-translational modifications in proteomics.

The open search strategy is not without limitations. Open searching can be computationally infeasible for large datasets without the use of the more efficient algorithms presented in Chapter Three. Aside from the computational costs, it does not recover all peptides all unmodified peptides, recovers only 50% of modified peptides (even for the most common modifications), and has difficulties identifying peptides with modifications near the peptide C-terminus (due to the unaccounted for shifted y-ions). Hence, we attempted to address these concerns and refine the open search strategy through MSFragger. In addition, we wanted to examine the diversity and abundances of modifications across experiments to understand if their impact on false positives is uniform across different experimental conditions or if the most common modifications are consistent enough to be targeted through simple variable modification searches. For these purposes, we performed large scale modification profiling of various shotgun proteomics experiments and also examined the presence of modified peptides in several proteomics applications.

4.2 Materials and methods

Datasets and Data Preparation

Six public datasets, all analyzed using the Thermo Scientific Q Exactive mass spectrometer was obtained and conversion of vendor .raw files to mzML was performed as previously described. Three of the six was used for the large-scale profiling studies: a HEK 293 dataset [65] (PXD001468, 1.12 million MS/MS spectra), a HeLa dataset [99] (HeLa proteome profiles from PXD000612, 2.8 million MS/MS spectra), and a triple-negative breast cancer (TNBC) dataset [66] (PRDB004167, 19.6 million MS/MS spectra). A clinical breast cancer dataset [102] (PXD000815, 34.3 million MS/MS spectra) was used for the SILAC analysis. A large scale study involving 5,188 LC-MS/MS runs [103] (raw data obtained from authors, 64.7 million MS/MS spectra) was used for the AP-MS analysis. Finally, a human RNA binding protein study

[104] (human RBP runs from PXD000513, 8834 MS/MS spectra) was used to demonstrate the utility of MSFragger in detecting RNA-crosslinked peptides.

Boosting unmodified peptides

MSFragger implements an unmodified peptide boosting feature. When invoked, PSMs that have an absolute value of the mass shift *dM* (defined as the difference between the theoretical and observed precursor peptide mass) less than the true precursor tolerance threshold specified by the search parameters ‘precursor_true_tolerance’ / ‘precursor_true_units’ are placed into a different scoring heap that only contains such unmodified peptides. After the calculation of expectations for all PSMs in both the regular and unmodified scoring heap, a ranking expectation is generated for all PSMs. For entries in the regular scoring heap, containing both modified and unmodified PSMs, the ranking expectation is the same as the computed expectation. The ranking expectation for entries in the unmodified peptides heap are modified based on the specified search parameters (multiplied by the specified expectation boost or an arbitrary small value for those that pass the ‘zero_bin_accept_expect’ expectation) and recorded as the ranking expectation. All PSMs are then merged and ordered by their ranking expectations prior to results reporting. It is important to note that the original expectations are reported rather than the ranking expectation.

Complementary ions for the recovery of C-terminal modifications

The addition of complementary ions follows the basic spectra pre-processing described previously [105,106]. The top N observed fragment ions, as specified by the ‘add_topN_complementary’ ions parameter are selected and are assumed to be either a singly charged y-ion for all spectra and a doubly charged y-ions for spectra with an identified charge state of 3+ or higher. The m/z of the complementary singly charged b-ion is then calculated from the calculated neutral mass of the assumed y-ion and the observed precursor mass. A complementary ion with this m/z and intensity equal to the y-ion from which it was derived is then inserted into the spectrum. Note that complementary ions are generated for both the singly charged and the doubly charged assumption of the observed fragment ion so that N complementary ions are inserted for spectra with charge state 2+ and 2N complementary ions are inserted for spectra with charge state 3+ or higher. These modified experimental spectra are then subjected to open database searching. As the original experimental fragment ion (from which the complementary ions are generated) is retained in the spectrum, it is possible that a single experimental observation can be incorrectly interpreted as multiple fragmentation events. Future work involving the addition of complementary ions to the theoretical spectrum instead will eliminate this problem and improve localization of modifications.

MS1-based precursor mass correction and identification based calibration

Instrument recorded precursor mass values for MS/MS spectra can be inaccurate while repeated observations of a precursor in survey (MS1) scans can be highly precise. A supplementary tool was developed as part of the MSFragger pipeline that, for each MS/MS event, takes the recorded m/z and retention time, examines the corresponding space in MS1 scans, and extracts the nearest peak feature by tracing the mass in retention time. The m/z is then calculated as a weighted average (by intensity) of all peaks in the trace. The precursor m/z for each MS/MS event is then updated with this value. For certain MS/MS events in which it was not possible to reconstruct the associated peak feature, no changes to the recorded m/z are made. Following precursor mass correction, identification-based mass recalibration of the MS/MS run is performed. In order to compare modification profiles that are resolved at sub-ppm levels across disparate experiments and labs, this calibration step is critical as slight deviations can cause broadening of features in the profile and loss of power in recovering modifications. To perform this calibration, unmodified peptide identifications (filtered at 1% PSM level FDR using the PeptideProphet probability) with observed mass difference dM less than 20ppm are selected. As instrument bias may drift over time and varies across m/z, a two-dimensional calibration grid is constructed using a retention time width of 5 minutes and an m/z width of 200 m/z. For each unmodified peptide, the corresponding cell in the grid is found. A weighted ppm bias, based on the proximity to each point, is added to each of the four points corresponding to that cell. The weighted averages on the calibration grid are then used to adjust the precursor m/z for all observed MS/MS events in the run. The corrected and calibrated m/z values are then written to a calibration file that is incorporated in downstream analysis.

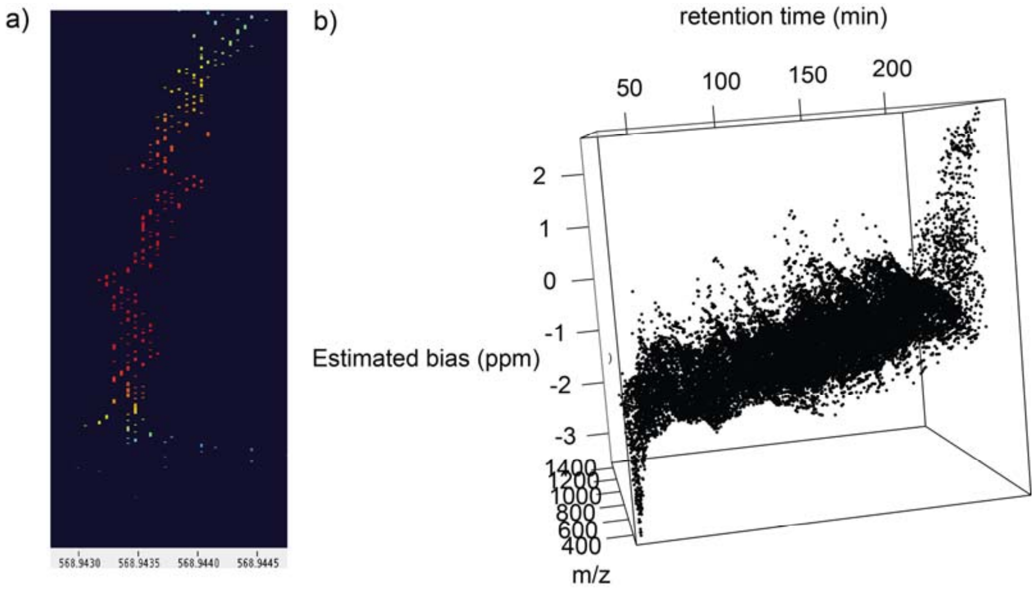


Figure 4-1 MS-1based precursor mass correction and identification based calibration.
(a) Visualization of a monoisotopic peptide ion peak using Batmass represented as a series of different observed m/z along retention time (vertical axis). (b) Calibration adjustment factors applied to each PSM plotted by their observed m/z and retention time.

Large scale profiling of chemical modification

Large scale profiling of chemical modifications was performed using the sequence database created from the human sequences of UniprotKB (Download date: 2015-10-09) appended with reversed protein sequences as decoys and common contaminants (cRAP proteins sequences from gpmDB and contaminants from MaxQuant). A precursor mass tolerance of 500 Da was used with fragment tolerance of 20 ppm. Isotopic error correction was disabled and common variable modifications of methionine oxidation and N-terminal acetylation were enabled. Carbamidomethylation was specified as a static modification. PSMs and peptides that contain modifications that were specified in our search parameters were not considered to have a mass shift for the tabulation of mass shifts. Fully tryptic digestion was specified allowing up to 1 missed cleavage. Complementary ions and boosting features were disabled and other MSFragger options were left as default.

MSFragger search results from each LC-MS/MS run were subjected to peptide validation as described above. Peptide probability was determined by the highest supporting PSM probability. Results for each experiment were aggregated and filtered at 1% peptide FDR. PSMs were separately filtered at 1% PSM FDR and only PSMs that passed both the 1% PSM FDR and 1% peptide FDR were retained for downstream analysis.

Modeling of observed modification profiles and detection of modification peaks

Normalized density profiles for each experiment were generated for comparison across different experiments. Corrected mass differences, with random noise on the order of +/- 5 μDa added to break ties, were binned using 0.0002 Da bins to form an initial counts histogram. These counts were then distributed to adjacent bins using the weights 0.23 (bin to left), 0.49 (same bin), 0.23 (bin to right) to smooth the histogram and improve the monotonicity of peak shapes. These histograms were then normalized by dividing each bin by the total number of spectra (in millions) acquired in the respective experiments. Averaging the counts in each bin generated an average profile of the three experiments. Mixture modeling of the average profile failed to precisely capture known modifications. Examination of the profile revealed peaks of varying broadness and further examination revealed the peak shape to be a complex function of the charge state and m/z of the underlying PSMs. Instead, a prominence based peak detection method was used that found features on the histogram by requiring that the peak prominence was at least 0.3 times that of the peak height. As known modifications were observed to have a peak width of approximately 0.004 Da (given current instrument accuracies and the correction/calibration method applied as described above), these features were ordered by the rise in density compared to the 0.003 Da flanking regions. It should be noted that some of the detected features (mass bins) could be artifacts of the peak picking algorithm, or may correspond to various combinations of multiple modifications.

Mass shift annotation using Unimod

The Unimod repository was downloaded (on 2016-04-22) in XML format and was parsed to extract modification names and mass shifts. Mass shifts associated with the addition or deletions of the twenty amino acids were appended to this list. Multiples of the mass difference between carbon-13 and carbon-12 were added as ‘First isotopic peak’ and ‘Second isotopic peak’ to account for isotopic peak picking errors. Entries that represent a single mass shift in this list were concatenated into a single entry so that a single text identifier represented each mass shift.

Annotation of the list of mass shifts proceeded in decreasing order of abundance. For each mass shift, the mass is queried against the described database of annotations with a mass tolerance of 0.002 Da. If a match is found, the mass shift is annotated with the entry from the database. If the mass shift cannot be matched to a single entry in the database, we attempt to compose multiple (up to 3) previously observed (in the order of annotation) mass shifts to account for compound modifications. If the mass shift remains unexplained, we add it to our list of annotations as a new un-annotated mass shift.

Localization of detected mass differences

For each PSM, including unmodified peptides, the observed mass difference is evaluated to see if it can be attributable to a modification of a specific site (position in the peptide). For each MS/MS run, the list of identified spectra (which includes the spectrum ID, peptide sequence, list of variably modified amino acids, and observed mass difference) is obtained from the MSFragger analysis pipeline, and the corresponding MS/MS spectra are extracted from the original mass spectrometry data file. The number of matched fragment ions is then re-computed using the same hyperscore function as originally done in MSFragger. The observed mass difference is iteratively placed on each amino acid, and for each position the spectrum similarity is computed to derive the number of matching fragment ions, and then the hyperscore. A PSM is called localizable if there is at least one position that generates a higher number of matched fragments than the rest. As there may be insufficient fragments to support an unambiguous localization in the peptide sequence, all positions that share the highest hyperscore are marked as a possible localization site. A PSM is called to be localized to the N-terminal if the localized positions form an uninterrupted stretch of amino acids from the N-terminal.

The localization results are then aggregated for each identified mass bin, and their localization characteristics examined. For each bin, the overall localization rate (the percentage of PSMs within that bin that are localizable), the N-terminal localization rate (the percentage of PSMs within that bin that are localizable and the localization is N-terminal), and the amino acid enrichment are computed. The amino acid enrichment is determined by first computing the amino acid composition of all peptides within the mass bin. Then, the number of localization sites attributable to each amino acid is summed across all localizable PSMs (for a PSM with multiple localization sites, each site gains a weight equal to 1 / number of localized sites). The total localization count for each amino acid is then normalized to form the localization rate. Amino acid enrichment is then determined by the ratio of localization rate to composition rate. It should be noted that while this metric is informative in many cases, it may be misleading in bins containing few PSMs or bins that are dominated by several abundant peptides that skews the counts and normalization factors.

Spectral similarity scores for modifications

For each modified PSM, we identify corresponding PSMs of the same charge state that identifies the same peptide but with a mass difference of less than 0.001 Da (indicating an unmodified peptide). We compute the average cosine similarity between the spectrum of the modified PSM and spectra corresponding to the unmodified peptide (if there are more than 50 such spectra, 50 are chosen at random). We then normalize for variations within unmodified spectra by dividing the average cosine similarity within the set of unmodified spectra to obtain a similarity score for the modified PSM. For each modification mass, its similarity score is determined by averaging the similarity scores calculated for each modified PSM within its mass tolerance.

Analysis of SILAC datasets

The breast cancer SILAC dataset was analyzed using the same search settings as the large-scale modification profiling described above with the exception that two variable modifications were added for the heavy labeled residues: 8.0142 Da at lysine and 10.00827 Da at arginine. Precursor mass correction/calibration and peptide validation were performed on each file and the aggregated files from the experiment were subjected to a 1% peptide and PSM FDR filter (each retained PSM passed 1% PSM FDR and matched a peptide that passed 1% peptide FDR). Each PSM in the resultant list was then examined for the presence of a heavy labeled residue (as determined by identification with a heavy labeled variable modification). Unlabeled PSMs were considered to have originated from the patient samples while labeled PSMs were considered to have originated from the super-SILAC mix.

Analysis of AP-MS dataset

Open search parameters for the AP-MS dataset were also similar to the settings used for large-scale modification profiling with one exception. As iodoacetamide treatment of samples was not used, no static modification was specified for cysteine. Each of the 5,188 runs was subjected to peptide validation and mass correction individually. FDR filtering was performed for each run individually, filtering the data at 1% FDR (at both peptide and PSM levels). Narrow window searches were performed using the same parameters with the exception of a 20 ppm precursor tolerance window and isotope selection errors of 0/1/2 was enabled.

For each LC-MS/MS run, all PSMs that were matched to a UniProt accession associated with the bait protein were considered to have originated from the bait protein (including any shared peptides). The number of unique sequences was determined by examining the set of unique peptides represented by the PSMs. Total counts for a particular bait protein across the replicates were determined by summing bait PSMs across the two replicates and determining the number of unique peptide sequences. Average fold change between narrow window and open searches was determined by linear regression in R.

Analysis of RNA-protein crosslink dataset

Open searching for the crosslinking dataset was performed similar to the large-scale modification profiling searches. The precursor mass window was enlarged to +/- 1000 Da to accommodate heavier crosslinked fragments. Carbamidomethylation was not specified as a fixed modification on cysteine. Comparison of results obtained by RNPxl and MSFragger was performed using the peptide sequence and mass difference. Identifications from RNPxl were translated into a peptide sequence and a total RNA-peptide mass. Identification from MSFragger was considered to be a match if it shared the same peptide sequence and had a total mass that differed from the RNPxl-based identification by no more than 0.05 Da.

4.3 Results

4.3.1 Refinement of the open search strategy using MSFragger

The development of MSFragger algorithm not only makes open searching practical, but also presents an opportunity to further investigate and refine this computational strategy. It is often assumed that the number of identified unique peptide sequences would be greatly reduced in open search compared to narrow window search due to the vastly expanded search space. However, our results using multiple search engines (Table 3-3) demonstrate that this is generally not the case. At the same time, it is true that not all unmodified, tryptic peptides found in narrow window search are found in open search. To see if those peptides can also be recovered, we implemented a boosting feature within MSFragger that preferentially ranks unmodified peptides over modified peptides when performing open search (see section 4.2). However, such a strategy, while implemented as an option in MSFragger, has not been found to significantly improve the results (Figure 4-2).

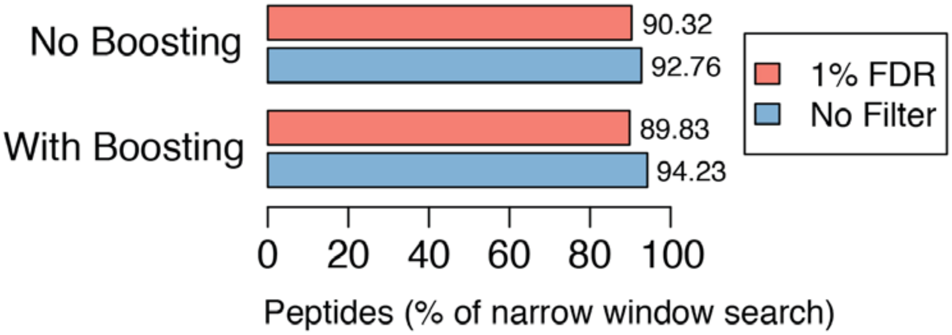


Figure 4-2 Preferential boosting of unmodified peptides fails to rescue missing peptides.
Boosting recovers a greater percentage of the peptides found in narrow window search prior to FDR filtering. Note that not all peptides identified in narrow window search are recovered in open search with the boosting option enabled due to the presence of a default peptide probability filter of 0.05 in PeptideProphet (disabling this filter using the –p0 option results in near 100% recovery). However, after controlling for FDR, boosting does not improve the peptide overlap between open and narrow window search.

In open searching, there is often a reduction in sensitivity for modified peptides containing common modifications that are specified as variable in narrow window searching (as open searching does not account for modified fragment ions). In order to address this issue, the speed of MSFragger allows us to specify variable modifications in conjunction with open searching.

We selected peptides identified with a single oxidized methionine in narrow window search and examined the proportion of such peptides observed in open search (explicitly as an oxidized methionine containing peptide or with a mass difference of +16 Da). In an open search without variable modifications, 45.4% of the peptides with oxidized methionine could be recovered in comparison to 87.5% for all peptides (Figure 4-3). Specifying oxidized methionine as a variable modification in our open search boosted the percentage of recovered oxidized methionine peptides to 88.8%, which is close to the 90.3% recovered for all peptides.

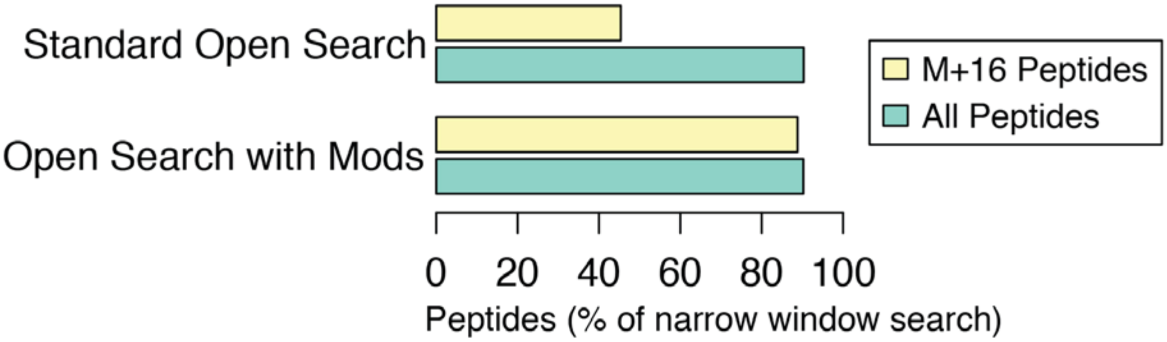


Figure 4-3 Decreased sensitivity for common modifications in open searching can be overcome by specifying variable modifications.
Standard open searches tend to identify far fewer peptides modified with common modifications than narrow window searching specifying those modifications as variable modifications. This is due to decreased sensitivity when the shifted ions are no longer matched in open search. For the most abundant chemical modifications, this can result in a significant decrease in overall counts. The speed of MSFragger allows variable modifications to be specified in conjunction with open searching. Examining peptides with oxidized methionine reveals that standard open search recovers only 45.37% of the peptides originally identified with oxidized methionine in narrow window searching (with variably oxidized methionine). Specifying oxidized methionine as a variable modification in open search brings that percentage to 88.81%, close to the overall overlap in peptide identifications between narrow window and open searches.

One apparent weakness of the open search strategy, compared to other strategies for blind PTM analysis such as spectrum alignment, is that it only considers unmodified fragment ions in scoring. Thus, C-terminal modifications are more difficult to detect using open searching as most y-ions, which are the most abundant and commonly observed in CID/HCD fragmentation, are shifted by the modification mass. Under the assumption that the most intense fragment ions are shifted y-ions, generating complementary ions [76,105] using the experimental precursor mass would yield unmodified b-ions that can be detected in open searches. We tested this hypothesis by inserting 10, 20, and 30 complementary ions in the experimental spectrum as a preprocessing step prior to searching and benchmarked this process using peptides with a single oxidized methionine. The addition of complementary ions slightly decreased the identification rates of peptides with modifications near the peptide N-terminus but greatly increased identification rates of peptides with modifications near the C-terminus (Figure 4-4). For peptides with an oxidized methionine just upstream of the tryptic cleavage site, the number of identified peptides increased by 48% when 20 complementary ions were added. As the number of complementary ion increased, the overall identification rates decreased due to the addition of noise in the spectrum. In our experience, the insertion of 20 complementary ions is optimal for detection of C-terminal modifications using the open search strategy. Because the overall improvement in the number of identifications (taking into consideration the unmodified peptides as well) when using the complementary ions was not significant, we elected not to use this option for the majority of analyses presented in this work. A more efficient strategy to account for shifted fragments would be to add complementary ions to theoretical spectra via extension of the fragment ion indexing scheme, which we plan will pursue in future work.

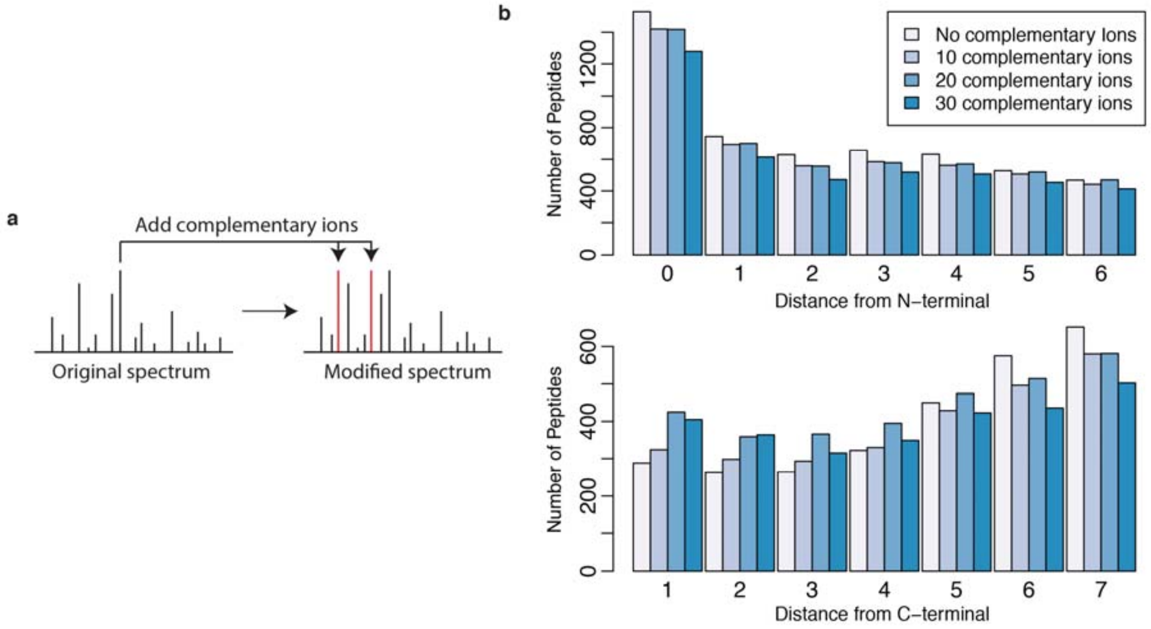
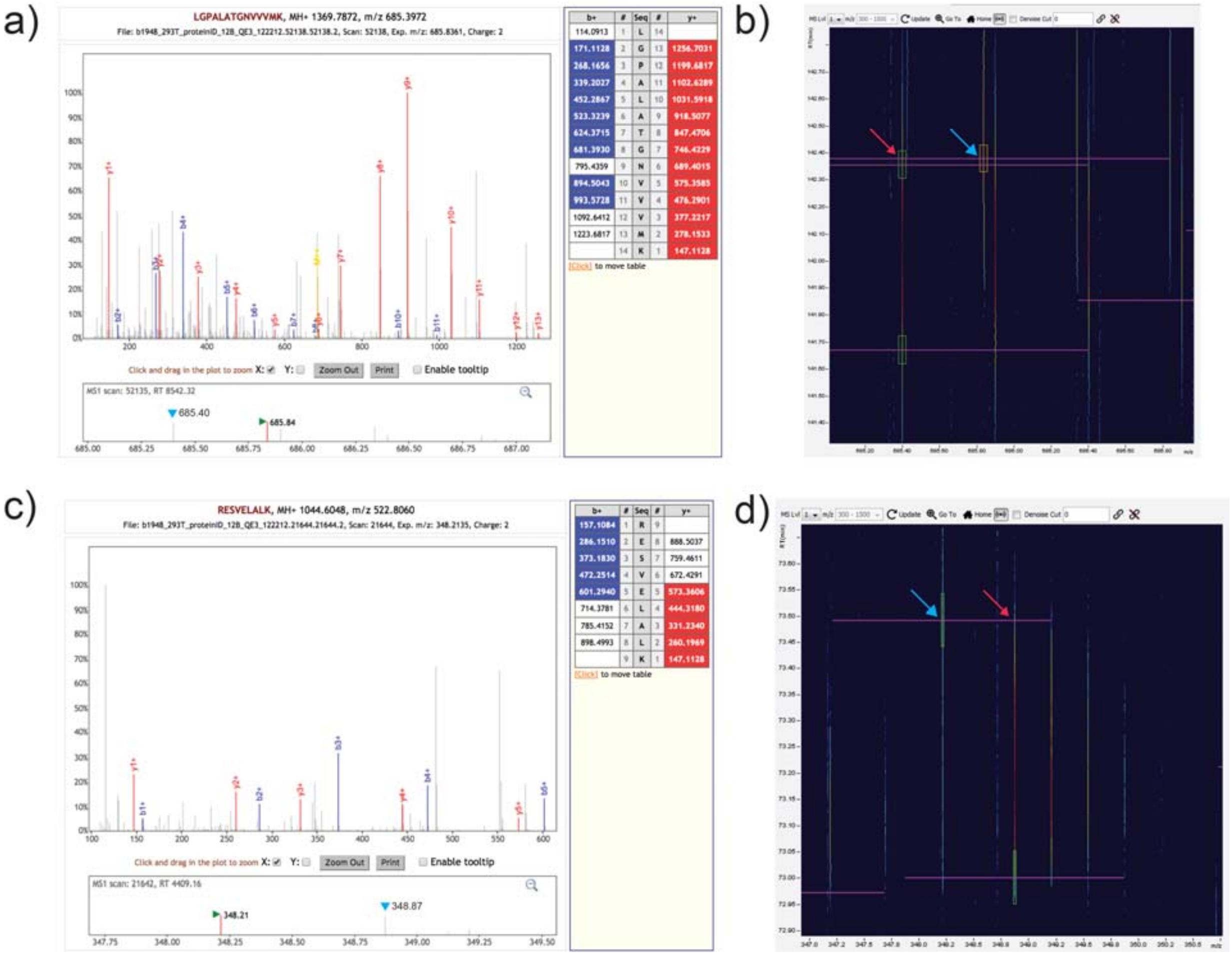


Figure 4-4 Complementary ions aid recovery of peptides with modifications near peptide C-terminus. (a) High intensity fragment ions are selected from the experimental spectrum and are assumed to be modified y-ions. Complementary ions based on the experimental precursor mass are inserted to form a modified spectrum that is subjected to open searching. (b) Evaluation of complementary ions using peptides containing a single oxidized methionine. 10, 20, and 30 complementary ions were inserted into each experimental spectrum and the counts of identified peptides were ordered by the distance of their oxidation site to the N or C-terminus. The addition of complementary ions decreased the number of identifications for peptides with oxidation near the N-terminus but greatly increased identification rates for peptides with oxidation near the C-terminus. For peptides with an oxidized methionine upstream of the tryptic cleavage site, the number of identified peptides increased by 48% when 20 complementary ions were added. The addition of more than 20 complementary ions was not found to be beneficial.

The problem of co-isolating multiple co-eluting peptides and the resulting chimera MS/MS spectra is well established [20,107] and manifests itself in unique ways in open searches. When a co-fragmented peptide is identified with a higher score, an artefactual (not attributed to any modification) mass shift is produced that can either be small (within several Daltons) or large (hundreds of Daltons) depending on whether the co-fragmented peptide ions are of the same or different charge state, respectively. Such cases can be identified using linked MS1 and MS/MS spectral viewers (Figure 4-5 a,c), and further evaluated using tools such as BatMass [91] (Figure 4-5 b,d). While the number of such cases is small, in future work chimeric spectra can be dealt with more accurately in open searches via MS1 feature detection of co-isolated peptides [37] within MSFragger or using external tools [108].

4.3.2 Large-scale profiling of unlabeled shotgun proteomics experiments

MSFragger's ultrafast performance enables comprehensive profiling of chemical and biological modifications across multiple large-scale proteomics datasets. To demonstrate this, we probed three large proteome-wide studies using open searches and compared their modification profiles. In addition to the HEK 293 dataset used to benchmark MSFragger, a HeLa proteome dataset [99], along with a dataset consisting of various triple-negative breast cancer (TNBC) cell lines and tissues [66] were used (see section 4.2). We additionally implemented a supplementary tool for MS1-based correction of precursor masses followed by identification-based mass recalibration. This allowed us to achieve sub-ppm mass accuracy and improved our ability to delineate modifications having close masses across disparate experiments and labs (Figure 4-6). The list of 500 most abundance mass shifts (excluding modifications specified as variable modifications in the search) is shown in Table C-2. We confirmed that in all datasets FDR estimates for modified peptides were well controlled and not inflated compared to unmodified peptides. For example, in HEK 293 dataset, peptide-level FDR was 0.18%, 0.11%, and 0.11% for peptides with top 500 most abundant mass shifts, top 100 mass most abundant shifts, and for unmodified peptides, respectively (FDRs computed separately for 500 most abundant mass shifts are shown in Table C-2).



LGPALATGNVVVMK, MH+ 1369.7872, m/z 685.3972

File: b1948_293T_proteinID_12B_QE3_122212.52138.52138.2, Scan: 52138, Exp. m/z: 685.8361, Charge: 2

b+	#	Seq	#	y+
114.0913	1	L	14	
171.1128	2	G	13	1256.7031
268.1656	3	P	12	1199.6817
339.2027	4	A	11	1102.6289
452.2867	5	L	10	1031.5918
523.3239	6	A	9	918.5077
624.3715	7	T	8	847.4706
681.3930	8	G	7	746.4229
795.4359	9	N	6	689.4015
894.5043	10	V	5	575.3585
993.5728	11	V	4	476.2901
1092.6412	12	V	3	377.2217
1223.6817	13	M	2	278.1533
	14	K	1	147.1128

[Click] to move table

Click and drag in the plot to zoom X: ☑ Y: □ Zoom Out Print □ Enable tooltip

MS1 scan: 52135, RT 8542.32

RESVELALK, MH+ 1044.6048, m/z 522.8060

File: b1948_293T_proteinID_12B_QE3_122212.21644.21644.2, Scan: 21644, Exp. m/z: 348.2135, Charge: 2

b+	#	Seq	#	y+
157.1084	1	R	9	
286.1510	2	E	8	888.5037
373.1830	3	S	7	759.4611
472.2514	4	V	6	672.4291
601.2940	5	E	5	573.3606
714.3781	6	L	4	444.3180
785.4152	7	A	3	331.2340
898.4993	8	L	2	260.1969
	9	K	1	147.1128

[Click] to move table

Click and drag in the plot to zoom X: ☑ Y: □ Zoom Out Print □ Enable tooltip

MS1 scan: 21642, RT 4409.16

Figure 4-5 Co-isolation of co-eluting precursors can result in mass differences that are not due to chemical modifications. (a) A MS/MS event was triggered at m/z 685.84 (green arrow) resulting in the identification of the peptide LGPALATGNVVVMK with a mass difference of 0.878. The parent survey scan reveals a co-eluting precursor with m/z 685.40 (cyan arrow). The difference in m/z at charge 2+ matches the observed mass difference suggesting that the co-eluting precursor is identified instead of the target precursor in this chimeric spectrum. (b) BatMass visualization of the MS/MS event described in (a) with MS/MS isolations marked by the purple line segments. The cyan arrow indicates the monoisotopic peak of the target precursor while the red arrow indicates the monoisotopic peak of the identified precursor. (c) The peptide RESVELALK was identified with a mass difference of -349.185 at m/z 348.21 (green arrow). Parent survey scan reveals a co-eluting precursor with m/z 348.87 (cyan arrow). While the target precursor ion is of charge 2+, the co-eluting precursor is of charge 3+, which transforms this 0.66 difference in m/z between these co-eluting precursors into the observed mass difference of -349.185. (d) Similar BatMass visualization of the MS/MS event described in (c). Note how the isolation window of the charge 2+ target precursor (cyan) crosses the monoisotopic peak of the charge 3+ co-eluting precursor (red).

We first interrogated several common chemical modifications (Figure 4-7a). Although the localization profiles were largely concordant (Figure C-1), their normalized abundances (modification rates) across the datasets were quite dissimilar. For example, the rate of phosphorylation in the HeLa dataset was over 14 times than that in the TNBC dataset.

Furthermore, some of these modifications were found on amino acids that are generally not considered in traditional workflows, such as tryptophan oxidation.

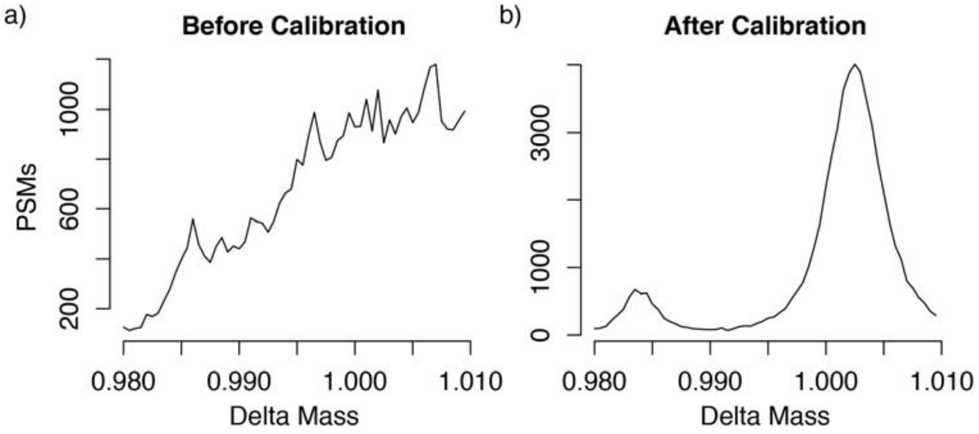


Figure 4-6 MS1-based correction of precursor masses and identification-based calibration helps delineate modifications in close mass proximity.
Identified number of PSMs with mass differences in the range of 0.98 Da to 1.01 Da from a single HEK293 LC-MS/MS run. Expected mass differences in this range are due to deamidation (with a delta mass of 0.984 Da) and C12/C13 error (with a delta mass of 1.003 Da). (a) Prior to correction a broad peak with no coherent shape is observed with a center around 1.005 Da. Knowledge of expected mass differences may lead to the calling of a peak near 0.986 Da. (b) Two cleanly resolved peaks are observed after mass correction. Expected peaks corresponding to deamidation and C12/C13 error are resolved with mean mass accurate to 1/1000 Da. The ability to determine such peaks from a single LC-MS/MS run demonstrates the accuracy of modern instruments and the power of our mass correction procedure.

We observed many highly abundant modifications that lacked annotations in Unimod and were unique (or of much greater abundance) to a particular dataset (Table C-2). To help decipher these unannotated (based on Unimod) modifications, we performed site localization analyses (section 4.2; Table C-1). For example, the HeLa dataset contained many peptides (over 23,000 PSMs) with a modification mass of 52.913 Da that were often localized to aspartic acid or glutamic acid, characteristic of metal ion adducts. This is likely to be iron displacing three protons (Unimod annotates ‘Replacement of 2 protons by iron’ modification only). We observed that many unannotated (in Unimod) mass shifts occurred on cysteines (Figure 4-7b). While some can be explained (e.g. 151.996 corresponding to carbamidomethylated DTT modification of cysteine), deducing the identities of unannotated mass shifts was outside the scope of this work.

For some modifications, we were unable to localize the mass shift on the peptide (Figure 4-7c). This suggests that there are few fragments that support the modification mass or that the detected modification mass is the result of a multiple modifications found on the same peptide. To investigate such cases, for each modification mass we computed a spectral similarity score between peptides containing that modification and their corresponding unmodified forms (see section 4.2; Figure 4-8). Most modifications possessed a similarity score between 0.4 and 0.6, including known modifications such as phosphorylation. However, we observed a large number of modifications (e.g. 3417 PSMs with mass shift 301.986 Da in HEK 293 dataset, 3068 PSMs with mass shift 284.126 Da in HeLa dataset) with similarity scores close to 1, indicating that spectra for many of the peptides with these modifications were largely unchanged from that of the unmodified peptide (Figure C-2). The lack of differences in the spectra and relative uniqueness to a particular dataset (Table C-2) suggests labile modifications that are specific to sample preparation protocol.

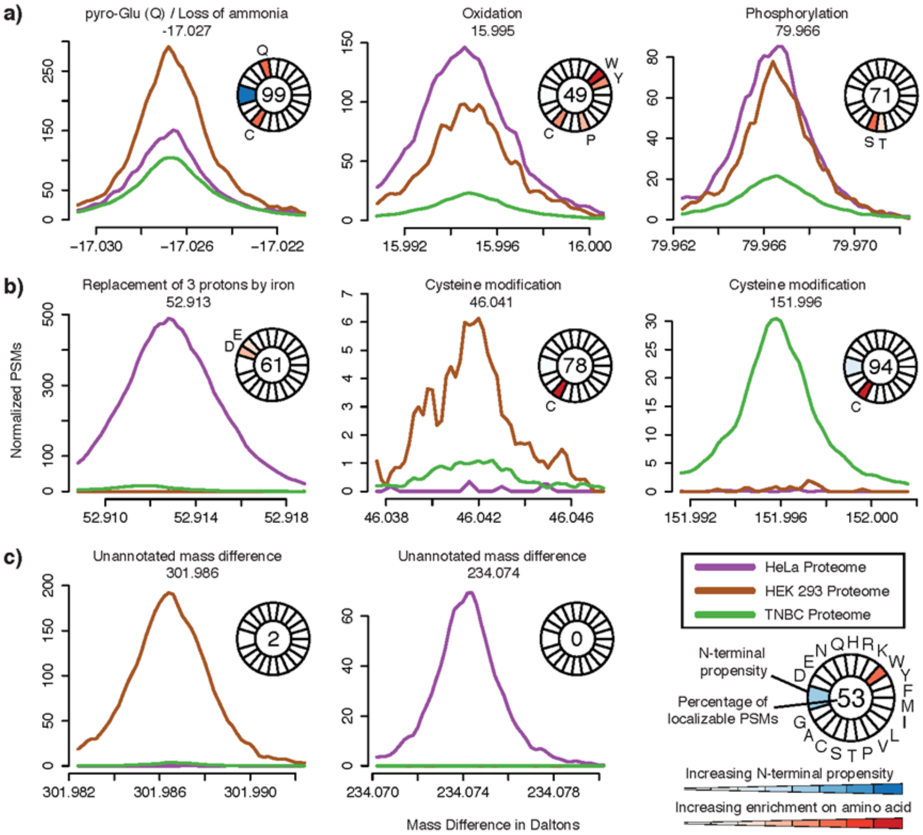


Figure 4-7 Modification profiles in large-scale HeLa, HEK293, and TNBC shotgun proteomics experiments.
(a) Examples of common modifications showing differences in modification rates. (b) Examples of abundant modifications that were unique to particular experiments. (c) Examples of abundant mass features where the mass difference could not be effectively localized.

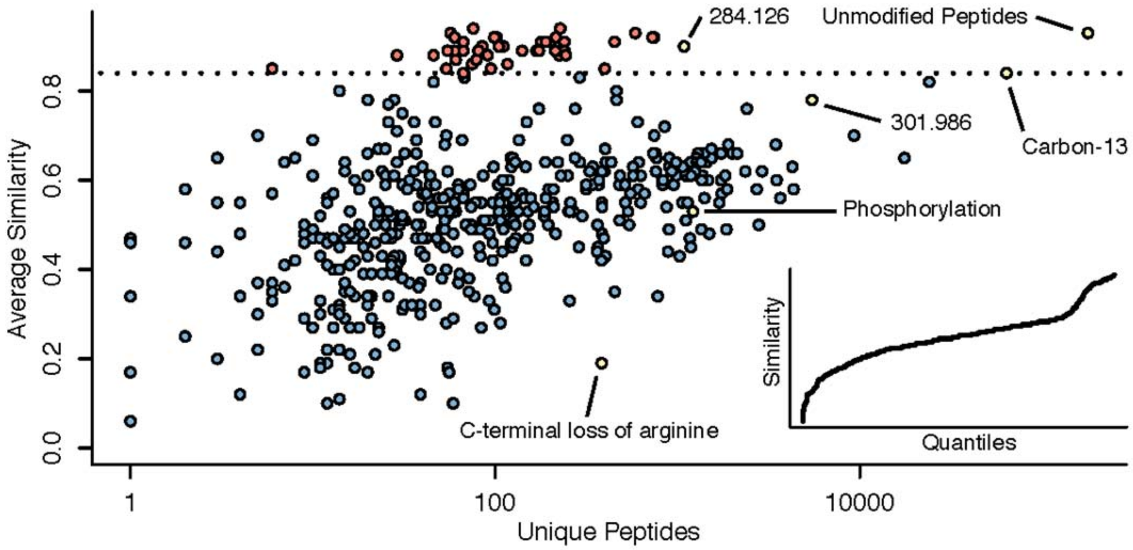


Figure 4-8 Open searching detects modified peptides containing labile modifications.
Spectral similarity scores for each mass bin were computed to capture the spectral similarity between a modified peptide and its unmodified counterpart. Spectra acquired from a ^{13}C ($^{12}C/^{13}C$) error are highly similar to those acquired from the unmodified form serving as a natural threshold (dotted line) of similarity significance between modifications with a higher similarity score (red) and those with a lower score (blue). Mass differences of interest are labeled and shown in yellow. Inset, quantile plot of similarity scores across modifications.

4.3.3 Modified peptides in various proteomics applications

MSFragger enables a wide range of analyses beyond interrogation of unlabeled proteomes. First, we are able to perform open searches using spectra from labeling-based experiments (e.g. SILAC, TMT, or iTRAQ) by specifying the labeled amino acids as a variable modification, thus allowing quantitative comparison of the modification states of proteins en masse. To test this, we examined a breast cancer dataset consisting of 442 LC-MS/MS runs representing 88 formalin-fixed paraffin-embedded (FFPE) patient samples that were analyzed together with a heavy labeled super-SILAC mix [102]. The open search (with variable SILAC modifications) of over 34 million MS/MS spectra from this dataset took less than three days. Examination of the modification profiles revealed a wide range of abundant modifications in these samples, as well as uncovered differences in modification abundances between the breast cancer samples and the super-SILAC mix, including a 30.011 Da mass shift that likely represents a methylol adduct which is characteristic of FFPE proteomes [109] (Figure 4-9a).

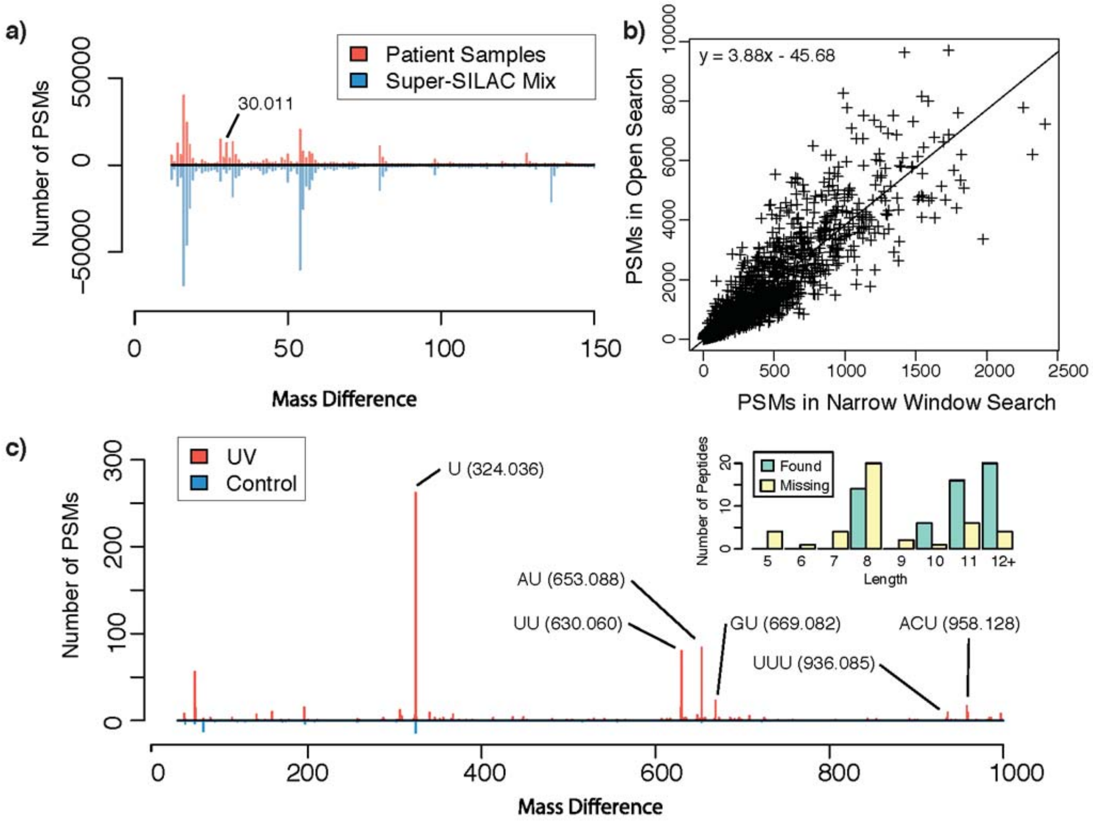


Figure 4-9 Application of MSFragger to diverse proteomics experiments.
(a) Comparison of a panel of breast tissue samples and a heavy-labeled super-SILAC mix, where SILAC-labeled amino acids were specified as variable modifications in conjunction with open searching. (b) Bait PSM counts identified in narrow-window and open searches in an AP–MS data set. (c) Open searching of an RNA–protein cross-linking data set. Prominent mass differences corresponding to RNA fragments are labeled. Inset, length of cross-linked peptides recovered by MSFragger.

Next, we applied MSFragger to a large-scale protein interaction study using an affinity purification mass spectrometry (AP-MS) experimental workflow that consisted of 2,594 baits analyzed in technical duplicates [103]. We reasoned that lowered sample complexity in AP-MS experiments provides an opportunity to examine in-depth the modification state of enriched proteins, most notably the proteins used as baits. We performed both narrow window searches and open searches on over 64.6 million MS/MS spectra across 5,188 LC-MS/MS runs. Open search increased the total number of PSMs by 32%, similar to the increases observed for data from whole cell lysates. For the bait proteins, however, the number of identified PSMs increased, on average, by almost 300% (Figure 4-9b).

For some bait proteins the increase in the number of identified peptide ions and total PSMs was astonishing. For example, the mitochondrial persulfide dioxygenase protein ETHE1 - a key member of the sulfur oxidation pathway that is itself involved in reactive oxidation of cysteine residues [110] - was identified by 48 and 2474 peptide ions in narrow window and open search, respectively. A large fraction of this increase for was attributed to cysteine modifications. When we subjected the top 100 bait proteins having the largest increase in the number of identified peptide ions to functional enrichment analysis using DAVID [111], the top enriched GO: Biological Process category (p-value 0.00007) was ‘small molecule metabolic process’ containing 23 proteins from the selected list, including ETHE1 (Table 4-1). Proteins in this category are involved in catalyzing modification processes and small molecule adducts, which may be linked to significantly higher number of modifications observed on these proteins themselves. These results suggest that application of MSFragger to affinity purification experiments can provide insights into a wide array of modifications, including rare and low abundance ones, on highly enriched proteins. Furthermore, open searching may offer better accounting of protein abundances using spectral counts in AP-MS experiments and improve the quality of recovered interaction networks derived using interaction scoring tools [112,113].

Table 4-1
List of genes associated with 'small molecule metabolic process' that have a large increase in identified bait peptide ions

ID	Gene Name
APIP	APAF1 interacting protein(APIP)
ACAA1	acetyl-CoA acyltransferase 1(ACAA1)
ACADSB	acyl-CoA dehydrogenase, short/branched chain(ACADSB)
AHCY	adenosylhomocysteinase(AHCY)
AGMAT	agmatinase(AGMAT)
ADH5	alcohol dehydrogenase 5 (class III), chi polypeptide(ADH5)
AKR7A3	aldo-keto reductase family 7, member A3 (aflatoxin aldehyde reductase)(AKR7A3)
CKB	creatine kinase, brain(CKB)
CIAPIN1	cytokine induced apoptosis inhibitor 1(CIAPIN1)
ETHE1	ethylmalonic encephalopathy 1(ETHE1)
GCDH	glutaryl-CoA dehydrogenase(GCDH)
GAMT	guanidinoacetate N-methyltransferase(GAMT)
GUK1	guanylate kinase 1(GUK1)
HMOX2	heme oxygenase 2(HMOX2)
LGMN	legumain(LGMN)
MVK	mevalonate kinase(MVK)
NEU2	neuraminidase 2 (cytosolic sialidase)(NEU2)
NUP43	nucleoporin 43kDa(NUP43)
PNMT	phenylethanolamine N-methyltransferase(PNMT)
PGAM2	phosphoglycerate mutase 2(PGAM2)
RPE	ribulose-5-phosphate-3-epimerase(RPE)
SORD	sorbitol dehydrogenase(SORD)
UROS	uroporphyrinogen III synthase(UROS)

Finally, we applied MSFragger to a RNA-protein crosslinking study [104]. Computational analysis for such studies can be challenging due to the need to determine a priori a list of potential crosslinked products. As open search allows for the identification of peptides with unknown modifications, no such list is required. Using a 1,000 Da precursor mass window, we performed open search on a run comprising of human UV-crosslinked RNA-protein complexes and a control non-irradiated run. We observed highly visible mass shifts associated with peptides crosslinked to mono, di, and tri-nucleotides in the irradiated sample that were largely absent from the control sample (Figure 4-9c). We compared our results to that of the RNPxl computational strategy described in the original study and found that open search confidently identified 163 crosslinked species, compared to 189 reported by RNPxl, with 134 identifications in common. As expected, the open search strategy failed to identify some of the crosslinked species containing very short peptides due to an insufficient number of unmodified fragment ions (Figure 4-9c inset**)**. On the other hand, MSFragger identified 29 additional crosslinked species, most of which (all except 4) were from proteins containing other crosslinked peptides already identified by RNPxl. Furthermore, MSFragger also identified a number of modified peptides from various RNA-binding proteins (including some not identified by RNPxl) with mass shifts that approximate the RNA crosslinks. These peptides are likely crosslinked peptides that also contain some other chemical modification or adduct and are thus undetectable by the RNPxl strategy. Examples include the peptides YGRPPDSHHSR and SYGRPPPDVEGMTSLK from the protein SRSF2 (which was not identified by RNPxl despite identifying 5 other proteins from the SRSF family). This shows that MSFragger provides a simple but highly effective analysis workflow for identification of protein-RNA crosslinked peptides, and demonstrates the added insights gained through open searching in any experimental setup.

4.4 Discussion

The refinements to the open search strategy were met with varying success. We demonstrated that peptides with the most common modifications could be recovered by combining variable modifications with open searching, which is important for enabling the use of open searching in quantitative proteomics experiments using labeling technologies. Failure to recover unmodified peptides using a boosting option suggest that the missing peptides may be of borderline quality and could potentially be false positives that are eliminated by open searching. While the addition of complementary ions helped recover a number of modified peptides with modifications near the C-terminus, it did not increase the overall number of identified peptides. Directly searching for shifted fragment ions in the future should help increase the recovery of such modifications as well as the overall number of modified peptides.

The vast array of chemical and biological modifications adds a dimension to proteomics that is not fully explored in most studies. Open database searching, made practical using MSFragger can, in conjunction with existing workflows, simultaneously and comprehensively identify modified and unmodified peptide forms. The diversity of chemical modifications in different experiments is revealing in that we cannot simply use a predefined set of common modifications but that they vary in abundance or may be unique to particular experimental workflows or labs. Monitoring the rates of these common chemical modifications is also important for reproducibility in quantitative proteomics experiments, especially when relying on quantification of selected peptides as proxies for estimating abundance of their corresponding protein [114].

Given the fast growth of public repositories of MS data [100], MSFragger could be used to search for rare (including novel) biological modifications across many biological samples and experimental conditions, adding to the list of previous such discoveries [115]. This includes large-scale cancer proteomics studies [116,117] where using the open search strategy can potentially identify novel PTMs that are involved in cancer processes. Open searching could be advantageous for characterization of neoantigens and other endogenous peptides[71,118], many of which are present in modified forms. The comprehensive identification of modified peptides using the open search strategy not only reduces the number of false positives in any proteomics workflow, but also provides exciting opportunities for the study of post-translational modifications.

4.5 Data availability

Raw mass spectrometry files are available from public repositories as described. The processed data files supporting the findings of this study are available upon request.

4.6 Acknowledgments

Andy Kong performed the data analysis with assistance from Dr. Alexey Nesvizhskii. Dr. Alexey Nesvizhskii supervised the entire project.

CHAPTER V

FUTURE EXPLORATIONS
BEYOND THE REFERENCE PROTEOME

5.1 Conclusions

Mass spectrometry has emerged as the method of choice for high-throughput proteome analysis. Advances in instrumentation over the past decade have allowed us to collect data of such depth and quality that we can now observe the vast majority of proteins that are inferred by the protein coding genes in the genome. However, these advances have been accompanied by growing pains as the development of statistical and computational methods has not kept pace with that of instrumentation and they struggle with challenging tasks such as the identification of novel peptides in proteogenomics, notably in the control of false positive identifications. In this dissertation, I presented computational strategies that identified and reduced false positives for the purpose of improving the quality and sensitivity of proteogenomics studies.

In Chapter Two of the dissertation, I provided a direct strategy for studying false positives by using high scoring decoy identifications that are generated from the target-decoy approach. Using this strategy and existing tools in an automated fashion, I was able to determine the fractions of false positives that were produced due to ambiguous scoring functions, semi-tryptic peptides, and modified peptides. While the percentage may vary from dataset to dataset, in the triple negative breast cancer dataset examined, these three categories represent nearly 80% of all false positives. As we were able to automatically annotate them by using multiple search engines or expanded search spaces, this suggests that we could eliminate them by incorporating these strategies in regular peptide identification workflows. We also demonstrate that false positives due to modified peptides may be underestimated using the target-decoy strategy by violating the assumption that incorrect assignments match at equal rates to target and decoy sequences. Using both experimental and theoretically generated spectra, we demonstrate that there is a greater propensity for modified peptides to map to target sequences, likely due to spectra homology between the spectra of modified peptides and those of other unmodified target peptides.

Database searches using multiple search engines, semi-tryptic searches, and open searches (to identify modified peptides) all serve to provide additional information that can be used to reduce the number of false positives. However, the computational costs of performing all such searches are impractical. In Chapter Three, we presented two database search tools that are much more efficient than current database search tools, enabling comprehensive analyses on large datasets. The first tool, EGADS, uses GPUs to accelerate both in-silico digestion and similarity scoring and provided speedups of 30-40X over conventional tools in common search scenarios. The three different scoring kernels implemented in EGADS might allow it to serve as multiple search engines and resolve ambiguous identifications. The second tool, MSFragger, uses a fragment indexing algorithm to simultaneously score experimental spectrum against a range of theoretical spectra without the use of specialized hardware. For open searches, MSFragger is over 150X faster than existing tools making them feasible for identifying blind modifications in large datasets. Together, these tools eliminate the bottlenecks that might be cause by database searching, allowing for more comprehensive peptide identification and reduction of false positives.

In Chapter 4, we refined the open search strategy for identifying modified peptides and applied it to a large number of shotgun experiments as well as other proteomics applications. Refinements of the open search strategy demonstrated the compatibility of the strategy with variable modification searching allowing common modifications to be fully recovered and enabling open search to be applied to labeling based quantitative proteomics experiments. We also increased recovery of peptides with modifications near the C-terminus and discussed other nuances of the open search strategy. We profiled a number of shotgun experiments and discovered that rates of chemical and biological modifications are quite dissimilar between experiments with some possessing unique modifications. We also discovered a number of labile modifications that showed no evidence for the modification in their tandem mass spectra. Finally, we applied the open search strategy and identified modified peptides in a number of proteomics applications demonstrating their pervasiveness and the added information they provide.

5.2 Future Directions

The automated pipeline used to identify sources of high scoring decoys can be adapted to use the highly efficient database search tools developed to examine sources of false positives across other large datasets. The sampling depth of the experiment (from faster instruments or extensive fractionation) might increase the number of modified peptides sampled, leading to become a larger source of false positives. Low-resolution tandem mass spectra may also lower identification confidences for ambiguous identifications due to a more robust null distribution used in score calibration.

The integration of multiple search engines, search spaces, and modified peptides can be challenging. While the pipeline used to explain high scoring decoys proceeded in a multi-stage approach, applying a similar approach can result in the trapping of false positives at an early stage (i.e. multiple search engines agree on the same false positive identification when there is a much better semi-tryptic or modified peptide explanation). A statistical framework is needed to establish the prior distributions of each input score, including the abundance estimation of modifications, in order to calibrate the scores and select the best explanation for a particular spectrum. The integration of different search spaces have been examined to some extent [56] using existing tools [69] but it is unclear whether those tools can be effectively used to integrate these wildly different search spaces.

The apparent violation of the target-decoy assumption by modified peptides requires a much deeper examination in both regular shotgun experiments and proteogenomics studies. Further experiments are needed to eliminate the effect of chimeric spectra in the work based on observed experimental data. A simple strategy might involve re-searching experimental spectra with modified peptides that have the fragment ions of the identified peptide removed and rejecting them from downstream analysis if a high quality match is found (indicating a chimeric spectra). The extent of this violation can also be determined for different modifications in order to investigate whether they occur only for modifications that have masses corresponding to single amino acid substitutions. These estimations can then be used to infer the degree of FDR underestimation in current proteomics studies at the PSM, peptide, and protein levels.

While the goal of these tools is to improve the quality of proteogenomics studies, they can also be applied to existing repositories and datasets to identify false positives that have been included in the current observed proteome (proteins that have been "confirmed" by mass spectrometry but are actually false positives). A strategy might involve the re-analysis of all spectra present in repositories such as GPMDB and flag any suspicious spectrum. Proteins that are only supported by suspicious spectra can then be manually inspected and possibly removed from the list of identified proteins. Performing similar validation tasks for proteogenomics studies can be complicated by their heterogeneity and use of custom databases.

EGADS is limited in the design of its memory manager and cannot operate on large protein databases that might be found in meta-proteomics studies. Changing the way the protein database is loaded onto GPU memory is a possibility but a simpler approach may be to simply partition the database into smaller segments and store partial search results for each spectrum and combine the search results after the final segment has been processed (similar to how MSFragger partitions the search space when there is insufficient memory to hold the entire fragment ion index). The application to meta-proteomics could be quite interesting due to EGADS's ability to rapidly process large sequence search spaces.

EGADS can also be used as a first pass tool for the automated detection and annotation of raw data that is submitted to a repository of mass spectrometry data. A non-specific search of all proteins can quickly identify the organism and enzymatic cleavage patterns as well as instrument mass accuracy and calibration. This set of information can be used to correct the data and determine search parameters in subsequent processing of the dataset, all without manual intervention and annotation – which can be error prone.

The algorithm behind MSFragger can be used to power a new class of peptide identification tools that can enable analysis not currently feasible due to the computational costs. This includes the direct searching of spliced peptides or fusion peptides by searching protein subsequences (using the shifted fragment index described below) without regards for the precursor mass and merging the results afterwards to “stitch” the two subsequences into a complete peptide. Other applications might involve large-scale spectra clustering (at a repository level) to identify spectra that have very different fragmentation patterns that might be indicative of a false positive.

More immediately, the use of a shifted fragment index (indexing ions based on the difference between their mass and their precursor masses) will allow the searching of mass shifted fragments and improve the ability to identify modified peptides using the open search strategy. Care must be taken to not double count experimental fragments for both the y-ion and y+ Δ-ion series. Concurrent searching of both indices would prevent this double counting but can be technically challenging to implement. The use of this modified algorithm for identifying modified peptides should be compared to the conventional open search strategy to establish differences in identification rates and to determine if there are novel C-terminal modifications that can be identified using this strategy that cannot be identified using the conventional open search strategy.

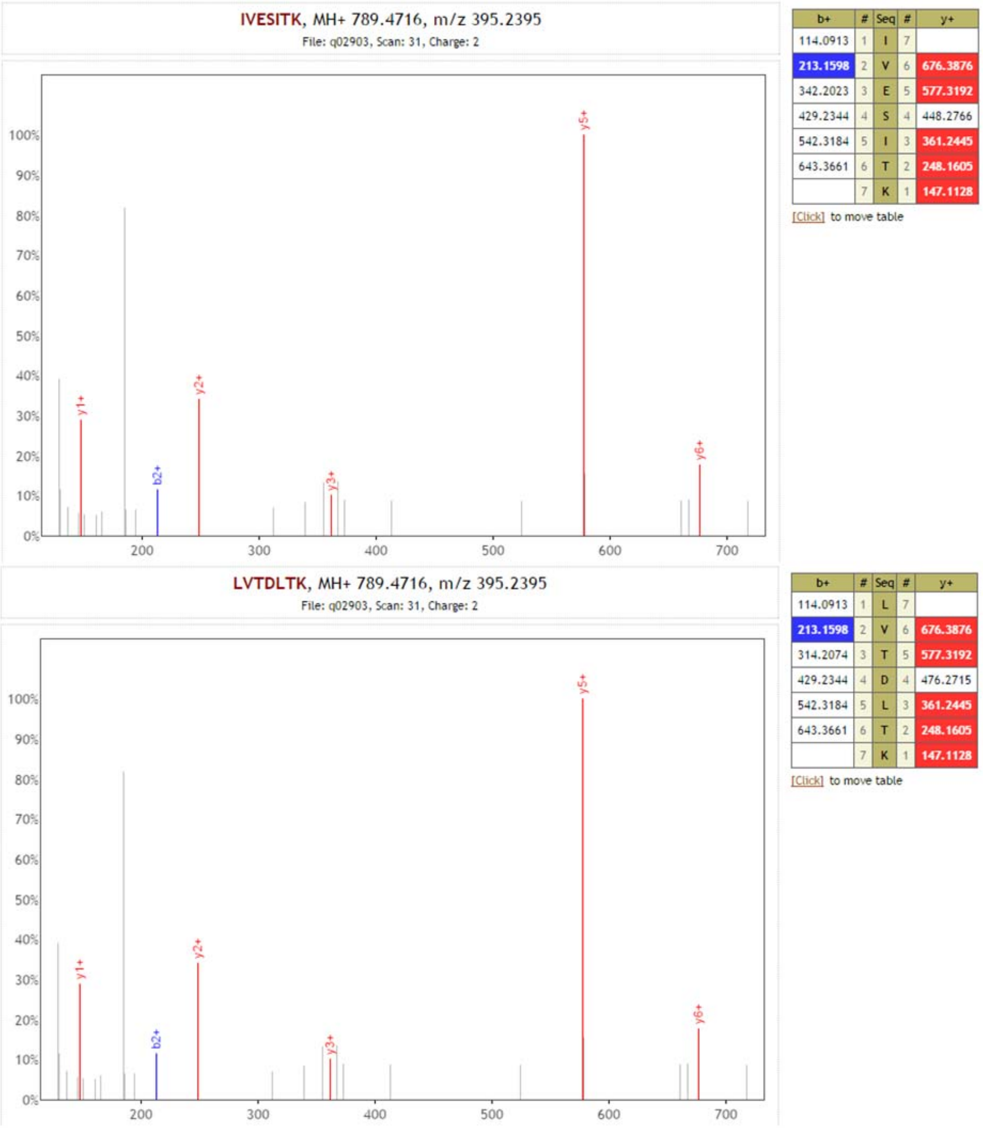
The use of the fragment ion-indexing algorithm can also be used to power direct searches against collections of spectra at a repository scale in real time, enabling researchers to establish mass spectrometry evidence for a particular peptide or modification interactively. A prototype of this application has demonstrated that millions of spectra can be effectively searched in tens of milliseconds but the scalability to hundreds of millions or billions of spectra remains to be tested. Alternative explanations from a variety of searches (enzymatic, semi-tryptic, open) will need to be pre-computed for each spectrum in the database to establish whether the newly tested hypothesis is better than the others that have been already tested. Statistical frameworks will need to be established to encompass the multiple search spaces from very heterogeneous experiments (different instruments, mass accuracies, digestions etc.).

The ability to identify large numbers of modified peptides in large-scale proteomics experiments is exciting as it opens up many avenues of inquiry from both a methods development and biological perspective. The varying rates of chemical modifications could have consequences for quantitative proteomics. Investigating ways on how to select peptides that are unlikely to be modified or incorporating the abundances of modified peptides in abundance estimation could improve such experiments by reducing variance and thus increase statistical power. Technical replicates of the same sample (and the same or different labs) could serve as benchmarking datasets.

The study of biological modifications from open search results could be a boon for the proteomics field as a whole as they cannot be assayed by inexpensive sequencing technologies. Even for well-characterized PTMs that have established enrichment protocols, the large number of non-enriched experiments covering a wide range of tissues could add tissue specific knowledge of these PTMs. Rare (and novel) PTMs can also be studied using open searching but precautions must be taken to avoid errors from incorrect charge state assignment or co-fragmentation. These modifications can be compiled into a resource for community curation and follow up by biochemical studies.

Appendix A

EXAMPLES OF HIGH SCORING DECOYS



IVESITK, MH+ 789.4716, m/z 395.2395

File: q02903, Scan: 31, Charge: 2

b+	#	Seq	#	y+
114.0913	1	I	7	
213.1598	2	V	6	676.3876
342.2023	3	E	5	577.3192
429.2344	4	S	4	448.2766
542.3184	5	I	3	361.2445
643.3661	6	T	2	248.1605
	7	K	1	147.1128

[Click] to move table

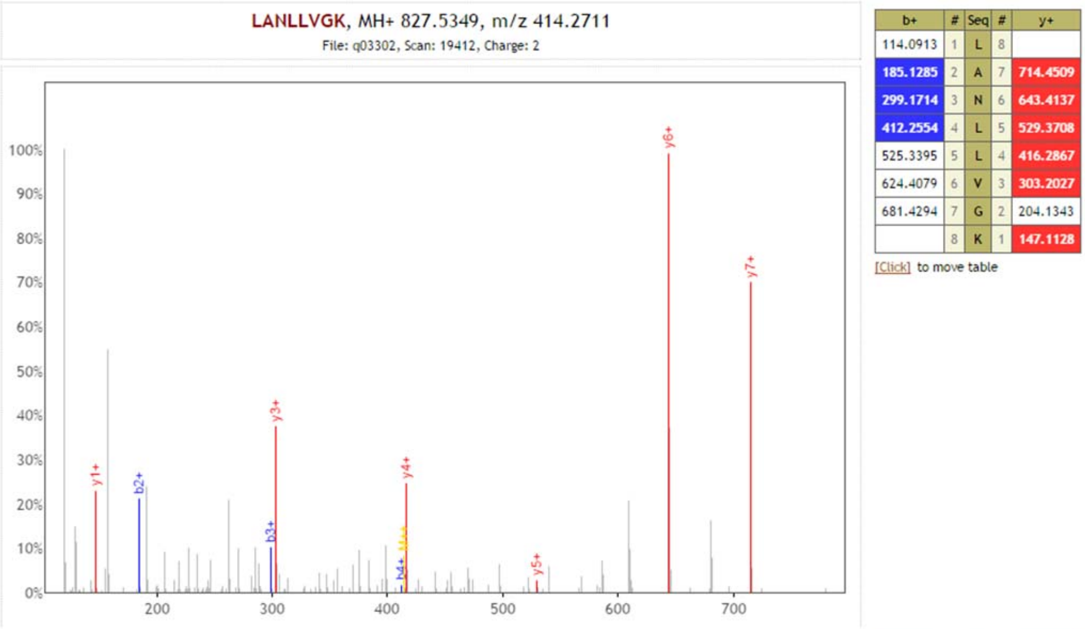
LVTDLTK, MH+ 789.4716, m/z 395.2395

File: q02903, Scan: 31, Charge: 2

b+	#	Seq	#	y+
114.0913	1	L	7	
213.1598	2	V	6	676.3876
314.2074	3	T	5	577.3192
429.2344	4	D	4	476.2715
542.3184	5	L	3	361.2445
643.3661	6	T	2	248.1605
	7	K	1	147.1128

[Click] to move table

Figure A-1 Example of high scoring decoy due to ambiguous scoring.
(Top) Decoy peptide identified with high confidence. (Bottom) Forward peptide identified with high confidence. The matched fragments are identical in mass indicating that both matches are of equal quality in the absence of a predicted fragmentation model.

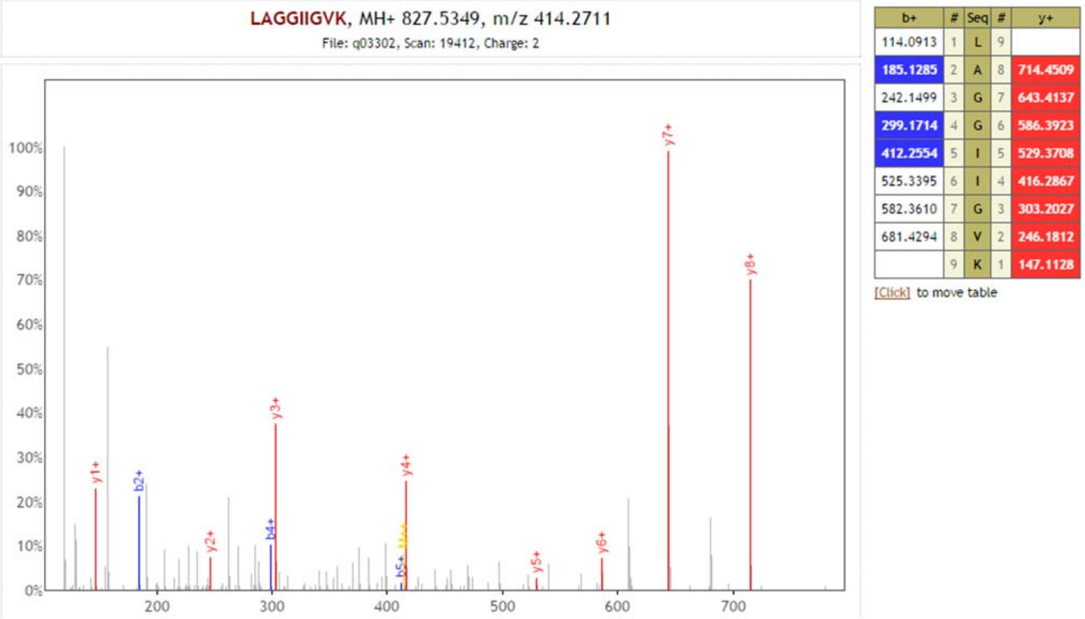


LANLLVGK, MH+ 827.5349, m/z 414.2711

File: q03302, Scan: 19412, Charge: 2

b+	#	Seq	#	y+
114.0913	1	L	8	
185.1285	2	A	7	714.4509
299.1714	3	N	6	643.4137
412.2554	4	L	5	529.3708
525.3395	5	L	4	416.2867
624.4079	6	V	3	303.2027
681.4294	7	G	2	204.1343
	8	K	1	147.1128

[Click] to move table



LAGGIIGVK, MH+ 827.5349, m/z 414.2711

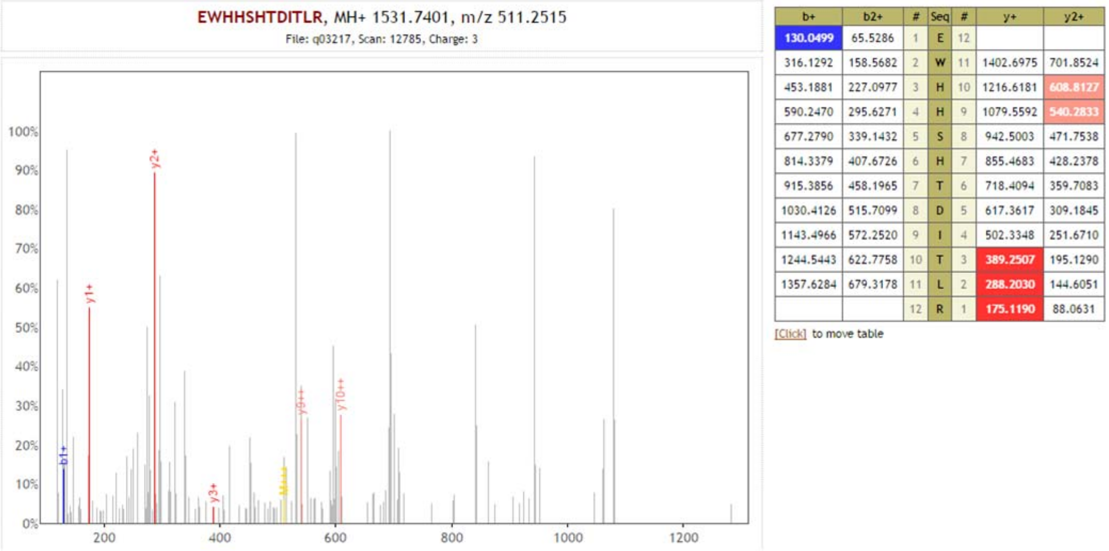
File: q03302, Scan: 19412, Charge: 2

b+	#	Seq	#	y+
114.0913	1	L	9	
185.1285	2	A	8	714.4509
242.1499	3	G	7	643.4137
299.1714	4	G	6	586.3923
412.2554	5	I	5	529.3708
525.3395	6	I	4	416.2867
582.3610	7	G	3	303.2027
681.4294	8	V	2	246.1812
	9	K	1	147.1128

[Click] to move table

Figure A-2 Example of high scoring decoy due to semi-tryptic peptide.

(Top) Decoy peptide identified with high confidence. (Bottom) Forward semi-tryptic peptide identified with high confidence.

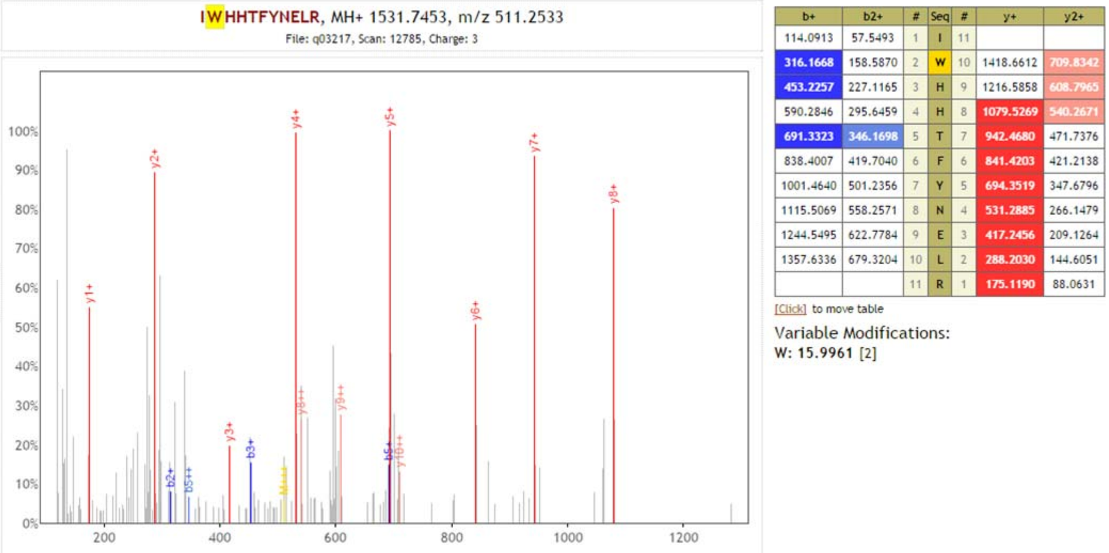


EWHHSHTDITLR, MH+ 1531.7401, m/z 511.2515

File: q03217, Scan: 12785, Charge: 3

b+	b2+	#	Seq	#	y+	y2+
130.0499	65.5286	1	E	12		
316.1292	158.5682	2	W	11	1402.6975	701.8524
453.1881	227.0977	3	H	10	1216.6181	608.8127
590.2470	295.6271	4	H	9	1079.5592	540.2833
677.2790	339.1432	5	S	8	942.5003	471.7538
814.3379	407.6726	6	H	7	855.4683	428.2378
915.3856	458.1965	7	T	6	718.4094	359.7083
1030.4126	515.7099	8	D	5	617.3617	309.1845
1143.4966	572.2520	9	I	4	502.3348	251.6710
1244.5443	622.7758	10	T	3	389.2507	195.1290
1357.6284	679.3178	11	L	2	288.2030	144.6051
		12	R	1	175.1190	88.0631

[Click] to move table



IWHHTFYNELR, MH+ 1531.7453, m/z 511.2533

File: q03217, Scan: 12785, Charge: 3

b+	b2+	#	Seq	#	y+	y2+
114.0913	57.5493	1	I	11		
316.1668	158.5870	2	W	10	1418.6612	709.8342
453.2257	227.1165	3	H	9	1216.5858	608.7965
590.2846	295.6459	4	H	8	1079.5269	540.2671
691.3323	346.1698	5	T	7	942.4680	471.7376
838.4007	419.7040	6	F	6	841.4203	421.2138
1001.4640	501.2356	7	Y	5	694.3519	347.6796
1115.5069	558.2571	8	N	4	531.2885	266.1479
1244.5495	622.7784	9	E	3	417.2456	209.1264
1357.6336	679.3204	10	L	2	288.2030	144.6051
		11	R	1	175.1190	88.0631

[Click] to move table

Variable Modifications:
W: 15.9961 [2]

Figure A-3 Example of high scoring decoy due to unaccounted for modification.
(Top) Decoy peptide identified with high confidence. (Bottom) Forward peptide identified with oxidation on tryptophan.

Appendix B

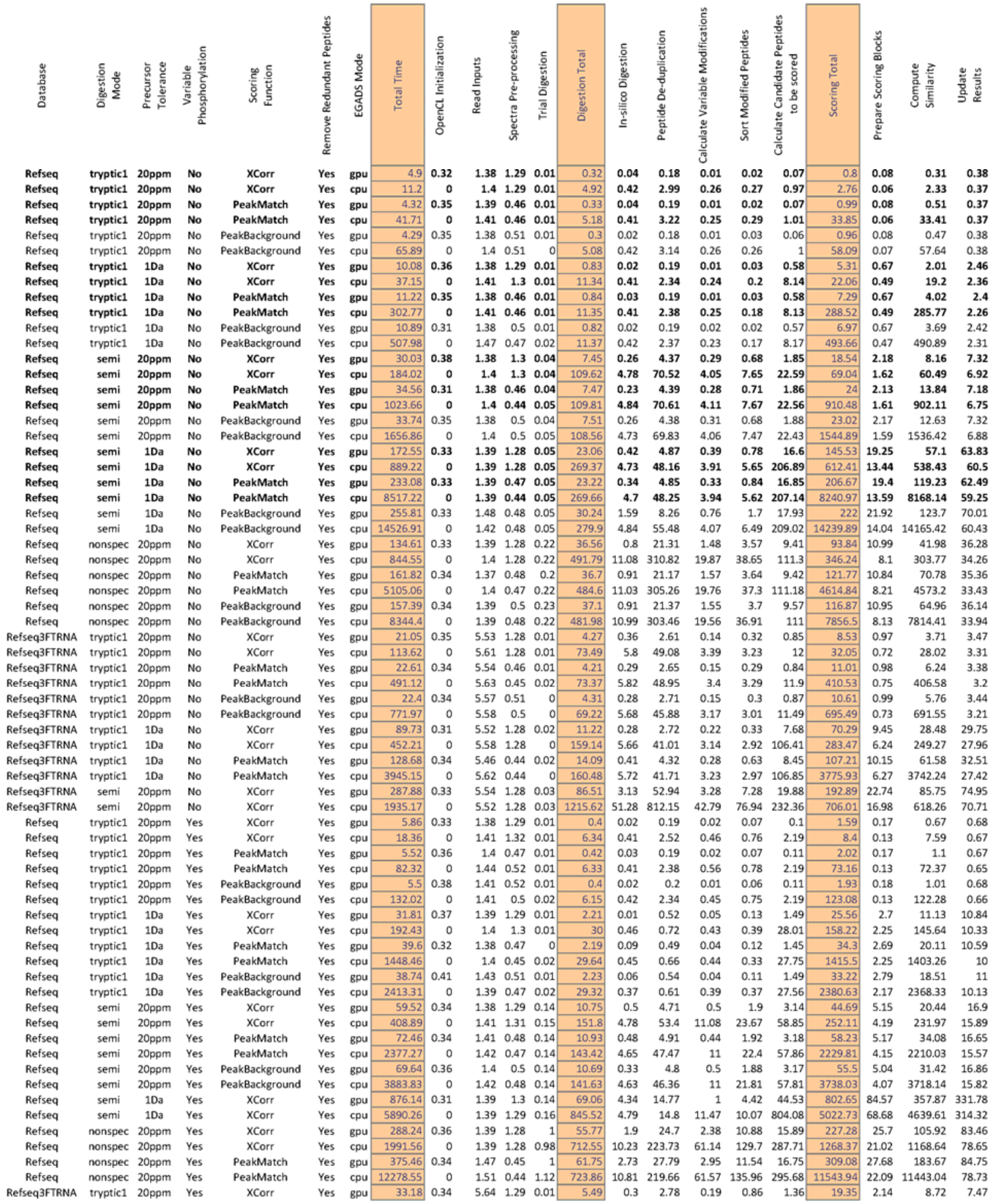
SUPPLEMENTARY MATERIALS FOR EFFICIENT DATABASE SEARCH TOOLS

Table B-1 Analysis times for a single file (b1906_293T_proteinID_01A_QE3_122212) in HEK293 dataset using different search engines.
All benchmarking was performed on a E3-1230v2 (4 cores with hyperthreading at 3.3 GHz) workstation unless otherwise noted using 8 threads. Breakdowns of indexing and search time provided where applicable. Times do not sum to overall search time due to input/output and other overhead.

Search Engine (search parameters)	Peptide Index Time	Fragment Index Time	Search Time	Overall Time
SEQUEST-HT* (500Da)	108.9 seconds	-	11.2 hours	11.3 hours
Comet (500Da)	-	-	-	13.6 hours
X!Tandem (500Da)	-	-	-	16.3 hours
MSFragger (500Da)	8.8 seconds	5.0 seconds	5.0 minutes	5.4 minutes
Comet (500Da with mods)	-	-	-	19.0 hours
X!Tandem (500Da with mods)	-	-	-	20.8 hours
MSFragger (500Da with mods)	11.7 seconds	12.1 seconds	6.9 minutes	7.4 minutes
SEQUEST-HT* (5ppm)	-	-	3.6 minutes	9.3 minutes
SEQUEST-HT* (100ppm)	-	-	3.7 minutes	9.5 minutes
Comet (100ppm)	-	-	-	101.3 seconds
X!Tandem (100ppm)	-	-	-	102.0 seconds
MSFragger (100ppm)	8.5 seconds	3.6 seconds	4.3 seconds	24.5 seconds
Comet (100ppm / low-res MS/MS)	-	-	-	64.1 seconds
X!Tandem (100ppm / low-res MS/MS)	-	-	-	122.4 seconds
MSFragger (100ppm / low-res MS/MS)	7.9 seconds	3.4 seconds	9.2 seconds	28.2 seconds
MODa (500Da, single-blind, fully tryptic)	-	-	-	150.3 minutes
MODa (500Da, single-blind, semi tryptic)	-	-	-	204.7 minutes
MSFragger (500Da)**	46.5 seconds	13.7 seconds	23.4 minutes	24.5 minutes
Tide (100Da)#	116.3 seconds	-	174.7 minutes	176.7 minutes
MSFragger (100Da)**	42.8 seconds	14.7 seconds	5.3 minutes	6.3 minutes

*performed on a 2xE5-2609v2 (2 processor, each with 4 cores at 2.5GHz) workstation, peptide indexing timed only for open search (PD uses cached index)
#Tide is unable to accommodate precursor tolerance windows larger than 100Da
**MSFragger was restricted to a single thread in comparisons against tools (MODa, Tide) that do not make use of multiple threads of execution

Table B-2 EGADS runtime across diverse search conditions.



Database	Digestion Mode	Precursor Tolerance	Variable Phosphorylation	Scoring Function	Remove Redundant Peptides	EGADS Mode	Total Time	OpenCL Initialization	Read Inputs	Spectra Pre-processing	Trial Digestion	Digestion Total	In-silico Digestion	Peptide De-duplication	Calculate Variable Modifications	Sort Modified Peptides	Calculate Candidate Peptides to be scored	Scoring Total	Prepare Scoring Blocks	Compute Similarity	Update Results
Refseq	tryptic1	20ppm	No	XCorr	Yes	gpu	4.9	0.32	1.38	1.29	0.01	0.32	0.04	0.18	0.01	0.02	0.07	0.8	0.08	0.31	0.38
Refseq	tryptic1	20ppm	No	XCorr	Yes	cpu	11.2	0	1.4	1.29	0.01	4.92	0.42	2.99	0.26	0.27	0.97	2.76	0.06	2.33	0.37
Refseq	tryptic1	20ppm	No	PeakMatch	Yes	gpu	4.32	0.35	1.39	0.46	0.01	0.33	0.04	0.19	0.01	0.02	0.07	0.99	0.08	0.51	0.37
Refseq	tryptic1	20ppm	No	PeakMatch	Yes	cpu	41.71	0	1.41	0.46	0.01	5.18	0.41	3.22	0.25	0.29	1.01	33.85	0.06	33.41	0.37
Refseq	tryptic1	20ppm	No	PeakBackground	Yes	gpu	4.29	0.35	1.38	0.51	0.01	0.3	0.02	0.18	0.01	0.03	0.06	0.96	0.08	0.47	0.38
Refseq	tryptic1	20ppm	No	PeakBackground	Yes	cpu	65.89	0	1.4	0.51	0	5.08	0.42	3.14	0.26	0.26	1	58.09	0.07	57.64	0.38
Refseq	tryptic1	1Da	No	XCorr	Yes	gpu	10.08	0.36	1.38	1.29	0.01	0.83	0.02	0.19	0.01	0.03	0.58	5.31	0.67	2.01	2.46
Refseq	tryptic1	1Da	No	XCorr	Yes	cpu	37.15	0	1.41	1.3	0.01	11.34	0.41	2.34	0.24	0.2	8.14	22.06	0.49	19.2	2.36
Refseq	tryptic1	1Da	No	PeakMatch	Yes	gpu	11.22	0.35	1.38	0.46	0.01	0.84	0.03	0.19	0.01	0.03	0.58	7.29	0.67	4.02	2.4
Refseq	tryptic1	1Da	No	PeakMatch	Yes	cpu	302.77	0	1.41	0.46	0.01	11.35	0.41	2.38	0.25	0.18	8.13	288.52	0.49	285.77	2.26
Refseq	tryptic1	1Da	No	PeakBackground	Yes	gpu	10.89	0.31	1.38	0.5	0.01	0.82	0.02	0.19	0.02	0.02	0.57	6.97	0.67	3.69	2.42
Refseq	tryptic1	1Da	No	PeakBackground	Yes	cpu	507.98	0	1.47	0.47	0.02	11.37	0.42	2.37	0.23	0.17	8.17	493.66	0.47	490.89	2.31
Refseq	semi	20ppm	No	XCorr	Yes	gpu	30.03	0.38	1.38	1.3	0.04	7.45	0.26	4.37	0.29	0.68	1.85	18.54	2.18	8.16	7.32
Refseq	semi	20ppm	No	XCorr	Yes	cpu	184.02	0	1.4	1.3	0.04	109.62	4.78	70.52	4.05	7.65	22.59	69.04	1.62	60.49	6.92
Refseq	semi	20ppm	No	PeakMatch	Yes	gpu	34.56	0.31	1.38	0.46	0.04	7.47	0.23	4.39	0.28	0.71	1.86	24	2.13	13.84	7.18
Refseq	semi	20ppm	No	PeakMatch	Yes	cpu	1023.66	0	1.4	0.44	0.05	109.81	4.84	70.61	4.11	7.67	22.56	910.48	1.61	902.11	6.75
Refseq	semi	20ppm	No	PeakBackground	Yes	gpu	33.74	0.35	1.38	0.5	0.04	7.51	0.26	4.38	0.31	0.68	1.88	23.02	2.17	12.63	7.32
Refseq	semi	20ppm	No	PeakBackground	Yes	cpu	1656.86	0	1.4	0.5	0.05	108.56	4.73	69.83	4.06	7.47	22.43	1544.89	1.59	1536.42	6.88
Refseq	semi	1Da	No	XCorr	Yes	gpu	172.55	0.33	1.39	1.28	0.05	23.06	0.42	4.87	0.39	0.78	16.6	145.53	19.25	57.1	63.83
Refseq	semi	1Da	No	XCorr	Yes	cpu	889.22	0	1.39	1.28	0.05	269.37	4.73	48.16	3.91	5.65	206.89	612.41	13.44	538.43	60.5
Refseq	semi	1Da	No	PeakMatch	Yes	gpu	233.08	0.33	1.39	0.47	0.05	23.22	0.34	4.85	0.33	0.84	16.85	206.67	19.4	119.23	62.49
Refseq	semi	1Da	No	PeakMatch	Yes	cpu	8517.22	0	1.39	0.44	0.05	269.66	4.7	48.25	3.94	5.62	207.14	8240.97	13.59	8168.14	59.25
Refseq	semi	1Da	No	PeakBackground	Yes	gpu	255.81	0.33	1.48	0.48	0.05	30.24	1.59	8.26	0.76	1.7	17.93	222	21.92	123.7	70.01
Refseq	semi	1Da	No	PeakBackground	Yes	cpu	14526.91	0	1.42	0.48	0.05	279.9	4.84	55.48	4.07	6.49	209.02	14239.89	14.04	14165.42	60.43
Refseq	nonspec	20ppm	No	XCorr	Yes	gpu	134.61	0.33	1.39	1.28	0.22	36.56	0.8	21.31	1.48	3.57	9.41	93.84	10.99	41.98	36.28
Refseq	nonspec	20ppm	No	XCorr	Yes	cpu	844.55	0	1.4	1.28	0.22	491.79	11.08	310.82	19.87	38.65	111.3	346.24	8.1	303.77	34.26
Refseq	nonspec	20ppm	No	PeakMatch	Yes	gpu	161.82	0.34	1.37	0.48	0.2	36.7	0.91	21.17	1.57	3.64	9.42	121.77	10.84	70.78	35.36
Refseq	nonspec	20ppm	No	PeakMatch	Yes	cpu	5105.06	0	1.4	0.47	0.22	484.6	11.03	305.26	19.76	37.3	111.18	4614.84	8.21	4573.2	33.43
Refseq	nonspec	20ppm	No	PeakBackground	Yes	gpu	157.39	0.34	1.39	0.5	0.23	37.1	0.91	21.37	1.55	3.7	9.57	116.87	10.95	64.96	36.14
Refseq	nonspec	20ppm	No	PeakBackground	Yes	cpu	8344.4	0	1.39	0.48	0.22	481.98	10.99	303.46	19.56	36.91	111	7856.5	8.13	7814.41	33.94
Refseq3FTRNA	tryptic1	20ppm	No	XCorr	Yes	gpu	21.05	0.35	5.53	1.28	0.01	4.27	0.36	2.61	0.14	0.32	0.85	8.53	0.97	3.71	3.47
Refseq3FTRNA	tryptic1	20ppm	No	XCorr	Yes	cpu	113.62	0	5.61	1.28	0.01	73.49	5.8	49.08	3.39	3.23	12	32.05	0.72	28.02	3.31
Refseq3FTRNA	tryptic1	20ppm	No	PeakMatch	Yes	gpu	22.61	0.34	5.54	0.46	0.01	4.21	0.29	2.65	0.15	0.29	0.84	11.01	0.98	6.24	3.38
Refseq3FTRNA	tryptic1	20ppm	No	PeakMatch	Yes	cpu	491.12	0	5.63	0.45	0.02	73.37	5.82	48.95	3.4	3.29	11.9	410.53	0.75	406.58	3.2
Refseq3FTRNA	tryptic1	20ppm	No	PeakBackground	Yes	gpu	22.4	0.34	5.57	0.51	0	4.31	0.28	2.71	0.15	0.3	0.87	10.61	0.99	5.76	3.44
Refseq3FTRNA	tryptic1	20ppm	No	PeakBackground	Yes	cpu	771.97	0	5.58	0.5	0	69.22	5.68	45.88	3.17	3.01	11.49	695.49	0.73	691.55	3.21
Refseq3FTRNA	tryptic1	1Da	No	XCorr	Yes	gpu	89.73	0.31	5.52	1.28	0.02	11.22	0.28	2.72	0.22	0.33	7.68	70.29	9.45	28.48	29.75
Refseq3FTRNA	tryptic1	1Da	No	XCorr	Yes	cpu	452.21	0	5.58	1.28	0	159.14	5.66	41.01	3.14	2.92	106.41	283.47	6.24	249.27	27.96
Refseq3FTRNA	tryptic1	1Da	No	PeakMatch	Yes	gpu	128.68	0.34	5.46	0.44	0.02	14.09	0.41	4.32	0.28	0.63	8.45	107.21	10.15	61.58	32.51
Refseq3FTRNA	tryptic1	1Da	No	PeakMatch	Yes	cpu	3945.15	0	5.62	0.44	0	160.48	5.72	41.71	3.23	2.97	106.85	3775.93	6.27	3742.24	27.42
Refseq3FTRNA	semi	20ppm	No	XCorr	Yes	gpu	287.88	0.33	5.54	1.28	0.03	86.51	3.13	52.94	3.28	7.28	19.88	192.89	22.74	85.75	74.95
Refseq3FTRNA	semi	20ppm	No	XCorr	Yes	cpu	1935.17	0	5.52	1.28	0.03	1215.62	51.28	812.15	42.79	76.94	232.36	706.01	16.98	618.26	70.71
Refseq	tryptic1	20ppm	Yes	XCorr	Yes	gpu	5.86	0.33	1.38	1.29	0.01	0.4	0.02	0.19	0.02	0.07	0.1	1.59	0.17	0.67	0.68
Refseq	tryptic1	20ppm	Yes	XCorr	Yes	cpu	18.36	0	1.41	1.32	0.01	6.34	0.41	2.52	0.46	0.76	2.19	8.4	0.13	7.59	0.67
Refseq	tryptic1	20ppm	Yes	PeakMatch	Yes	gpu	5.52	0.36	1.4	0.47	0.01	0.42	0.03	0.19	0.02	0.07	0.11	2.02	0.17	1.1	0.67
Refseq	tryptic1	20ppm	Yes	PeakMatch	Yes	cpu	82.32	0	1.44	0.52	0.01	6.33	0.41	2.38	0.56	0.78	2.19	73.16	0.13	72.37	0.65
Refseq	tryptic1	20ppm	Yes	PeakBackground	Yes	gpu	5.5	0.38	1.41	0.52	0.01	0.4	0.02	0.2	0.01	0.06	0.11	1.93	0.18	1.01	0.68
Refseq	tryptic1	20ppm	Yes	PeakBackground	Yes	cpu	132.02	0	1.41	0.5	0.02	6.15	0.42	2.34	0.45	0.75	2.19	123.08	0.13	122.28	0.66
Refseq	tryptic1	1Da	Yes	XCorr	Yes	gpu	31.81	0.37	1.39	1.29	0.01	2.21	0.01	0.52	0.05	0.13	1.49	25.56	2.7	11.13	10.84
Refseq	tryptic1	1Da	Yes	XCorr	Yes	cpu	192.43	0	1.4	1.3	0.01	30	0.46	0.72	0.43	0.39	28.01	158.22	2.25	145.64	10.33
Refseq	tryptic1	1Da	Yes	PeakMatch	Yes	gpu	39.6	0.32	1.38	0.47	0	2.19	0.09	0.49	0.04	0.12	1.45	34.3	2.69	20.11	10.59
Refseq	tryptic1	1Da	Yes	PeakMatch	Yes	cpu	1448.46	0	1.4	0.45	0.02	29.64	0.45	0.66	0.44	0.33	27.75	1415.5	2.25	1403.26	10
Refseq	tryptic1	1Da	Yes	PeakBackground	Yes	gpu	38.74	0.41	1.43	0.51	0.01	2.23	0.06	0.54	0.04	0.11	1.49	33.22	2.79	18.51	11
Refseq	tryptic1	1Da	Yes	PeakBackground	Yes	cpu	2413.31	0	1.39	0.47	0.02	29.32	0.37	0.61	0.39	0.37	27.56	2380.63	2.17	2368.33	10.13
Refseq	semi	20ppm	Yes	XCorr	Yes	gpu	59.52	0.34	1.38	1.29	0.14	10.75	0.5	4.71	0.5	1.9	3.14	44.69	5.15	20.44	16.9
Refseq	semi	20ppm	Yes	XCorr	Yes	cpu	408.89	0	1.41	1.31	0.15	151.8	4.78	53.4	11.08	23.67	58.85	252.11	4.19	231.97	15.89
Refseq	semi	20ppm	Yes	PeakMatch	Yes	gpu	72.46	0.34	1.41	0.48	0.14	10.93	0.48	4.91	0.44	1.92	3.18	58.23	5.17	34.08	16.65
Refseq	semi	20ppm	Yes	PeakMatch	Yes	cpu	2377.27	0	1.42	0.47	0.14	143.42	4.65	47.47	11	22.4	57.86	2229.81	4.15	2210.03	15.57
Refseq	semi	20ppm	Yes	PeakBackground	Yes	gpu	69.64	0.36	1.4	0.5	0.14	10.69	0.33	4.8	0.5	1.88	3.17	55.5	5.04	31.42	16.86
Refseq	semi	20ppm	Yes	PeakBackground	Yes	cpu	3883.83	0	1.42	0.48	0.14	141.63	4.63	46.36	11	21.81	57.81	3738.03	4.07	3718.14	15.82
Refseq	semi	1Da	Yes	XCorr	Yes	gpu	876.14	0.31	1.39	1.3	0.14	69.06	4.34	14.77	1	4.42	44.53	802.65	84.57	357.87	331.78
Refseq	semi	1Da	Yes	XCorr	Yes	cpu	5890.26	0	1.39	1.29	0.16	845.52	4.79	14.8	11.47	10.07	804.08	5022.73	68.68	4639.61	314.32
Refseq	nonspec	20ppm	Yes	XCorr	Yes	gpu	288.24	0.36	1.39	1.28	1	55.77	1.9	24.7	2.38	10.88	15.89	227.28	25.7	105.92	83.46
Refseq	nonspec	20ppm	Yes	XCorr	Yes	cpu	1991.56	0	1.39	1.28	0.98	712.55	10.23	223.73	61.14	129.7	287.71	1268.37	21.02	1168.64	78.65
Refseq	nonspec	20ppm	Yes	PeakMatch	Yes	gpu	375.46	0.34	1.47	0.45	1	61.75	2.73	27.79	2.95	11.54	16.75	309.08	27.68	183.67	84.75
Refseq	nonspec	20ppm	Yes	PeakMatch	Yes	cpu	12278.55	0	1.51	0.44	1.12	723.86	10.81	219.66	61.57	135.96	295.68	11543.94	22.09	11443.04	78.73
Refseq3FTRNA	tryptic1	20ppm	Yes	XCorr	Yes	gpu	33.18	0.34	5.64	1.29	0.01	5.49	0.3	2.78	0.19	0.86	1.36	19.35	2.14	8.72	7.47

Database	Digestion Mode	Precursor Tolerance	Variable Phosphorylation	Scoring Function	Remove Redundant Peptides	EGADS Mode	Total Time	OpenCL Initialization	Read Inputs	Spectra Pre-processing	Trial Digestion	Digestion Total	In-silico Digestion	Peptide De-duplication	Calculate Variable Modifications	Sort Modified Peptides	Calculate Candidate Peptides to be scored	Scoring Total	Prepare Scoring Blocks	Compute Similarity	Update Results
Refseq3FTRNA	tryptic1	20ppm	Yes	XCorr	Yes	cpu	208.63	0	5.63	1.3	0.02	90.64	5.66	40.2	6.27	10.41	28.1	109.59	1.81	100.76	7.02
Refseq3FTRNA	tryptic1	20ppm	Yes	PeakMatch	Yes	gpu	38.08	0.34	5.55	0.47	0.02	5.58	0.31	2.81	0.22	0.78	1.47	25.03	2.08	14.69	7.27
Refseq3FTRNA	tryptic1	20ppm	Yes	PeakMatch	Yes	cpu	1053.81	0	5.52	0.44	0.02	85.21	5.7	36.32	6.07	9.73	27.39	961.23	1.73	952.65	6.85
Refseq3FTRNA	tryptic1	20ppm	Yes	PeakBackground	Yes	gpu	37.38	0.39	5.55	0.48	0.02	5.66	0.34	2.84	0.22	0.81	1.45	24.16	2.19	13.45	7.52
Refseq3FTRNA	tryptic1	20ppm	Yes	PeakBackground	Yes	cpu	1708.48	0	5.52	0.47	0.02	85.03	5.69	36.23	5.91	9.75	27.45	1616.1	1.74	1607.35	7
Refseq3FTRNA	tryptic1	1Da	Yes	XCorr	Yes	gpu	468.64	0.31	5.54	1.28	0	34.39	0.88	8.1	0.61	1.88	22.92	425.69	44.3	188.84	177.45
Refseq3FTRNA	tryptic1	1Da	Yes	XCorr	Yes	cpu	3095.17	0	5.54	1.28	0.02	441.58	5.61	9.57	5.67	4.78	415.81	2635.78	36.58	2430.13	168.99
Refseq3FTRNA	semi	20ppm	Yes	XCorr	Yes	gpu	695.56	0.31	5.46	1.28	0.11	135.55	6.33	63.65	6.04	23.35	36.18	551.06	61.65	266.33	195.76
Refseq3FTRNA	semi	20ppm	Yes	XCorr	Yes	cpu	4392.52	0	5.47	1.25	0.13	1567.59	53.14	499.27	122.97	246.67	645.21	2803.43	47.67	2575.43	180.02
Refseq	tryptic1	20ppm	No	XCorr	No	gpu	5.39	0.32	1.38	1.29	0.01	0.22	0.03	0	0.02	0.04	0.13	1.3	0.14	0.48	0.62
Refseq	tryptic1	20ppm	No	XCorr	No	cpu	11.7	0	1.41	1.31	0.01	3.11	0.41	0	0.26	0.55	1.89	4.99	0.1	4.29	0.59
Refseq	tryptic1	20ppm	No	PeakMatch	No	gpu	5.02	0.34	1.41	0.47	0	0.23	0.03	0	0.03	0.04	0.13	1.69	0.15	0.86	0.62
Refseq	tryptic1	20ppm	No	PeakMatch	No	cpu	67.3	0	1.47	0.46	0.01	3.15	0.42	0	0.28	0.58	1.87	61.33	0.1	60.65	0.58
Refseq	tryptic1	20ppm	No	PeakBackground	No	gpu	4.99	0.33	1.44	0.51	0.01	0.21	0.02	0	0.02	0.04	0.13	1.63	0.15	0.8	0.62
Refseq	tryptic1	20ppm	No	PeakBackground	No	cpu	111.93	0	1.4	0.49	0	3.09	0.41	0	0.26	0.55	1.87	106.08	0.1	105.38	0.6
Refseq	tryptic1	1Da	No	XCorr	No	gpu	14.55	0.41	1.39	1.29	0.01	1.15	0.02	0	0.02	0.05	1.06	9.36	1.21	3.59	4.22
Refseq	tryptic1	1Da	No	XCorr	No	cpu	59.36	0	1.4	1.3	0.01	16.01	0.41	0	0.28	0.54	14.77	39.47	0.87	34.56	4.03
Refseq	tryptic1	1Da	No	PeakMatch	No	gpu	25.5	0.31	1.39	0.54	0	1.19	0.04	0	0.03	0.05	1.07	13	1.24	7.23	4.18
Refseq	tryptic1	1Da	No	PeakMatch	No	cpu	536.64	0	1.42	0.45	0.02	15.87	0.41	0	0.28	0.5	14.68	517.7	0.86	512.88	3.96
Refseq	tryptic1	1Da	No	PeakBackground	No	gpu	16.8	0.35	1.42	0.51	0.01	1.19	0.02	0	0.03	0.05	1.09	12.42	1.22	6.64	4.23
Refseq	tryptic1	1Da	No	PeakBackground	No	cpu	901.93	0	1.37	0.5	0	15.62	0.41	0	0.28	0.44	14.49	883.34	0.84	878.52	3.98
Refseq	semi	20ppm	No	XCorr	No	gpu	37	0.32	1.38	1.28	0.04	4.84	0.24	0	0.55	1.08	2.97	28.21	3.5	11.65	11.66
Refseq	semi	20ppm	No	XCorr	No	cpu	180.13	0	1.4	1.29	0.05	63.51	4.74	0	6.54	14.27	37.96	112.09	2.56	98.53	11
Refseq	semi	20ppm	No	PeakMatch	No	gpu	45.83	0.34	1.39	0.48	0.04	4.94	0.25	0	0.54	1.07	3.09	37.74	3.5	21.33	11.41
Refseq	semi	20ppm	No	PeakMatch	No	cpu	1521.57	0	1.39	0.47	0.05	61.2	4.73	0	6.41	13.23	36.83	1456.75	2.52	1443.49	10.75
Refseq	semi	20ppm	No	PeakBackground	No	gpu	44.04	0.33	1.39	0.5	0.05	4.9	0.31	0	0.45	1.12	3.01	35.96	3.47	19.5	11.58
Refseq	semi	20ppm	No	PeakBackground	No	cpu	2558	0	1.39	0.48	0.05	60.51	4.66	0	6.4	12.9	36.53	2493.89	2.47	2480.48	10.92
Refseq	semi	1Da	No	XCorr	No	gpu	259.91	0.33	1.39	1.29	0.05	28.77	0.53	0	0.61	1.17	26.46	227.06	30.4	88.27	99.81
Refseq	semi	1Da	No	XCorr	No	cpu	1333.15	0	1.44	1.31	0.03	356.45	4.77	0	6.63	11.65	333.38	966.9	21.38	850.72	94.8
Refseq	nonspec	20ppm	No	XCorr	No	gpu	161.48	0.34	1.39	1.3	0.23	23.08	0.99	0	2.42	5.49	14.19	134.17	16.56	56.51	54.05
Refseq	nonspec	20ppm	No	XCorr	No	cpu	800.92	0	1.4	1.28	0.23	274.97	11.13	0	30.31	63.37	170.11	517.98	11.86	455.49	50.59
Refseq3FTRNA	tryptic1	20ppm	No	XCorr	No	gpu	27.43	0.34	5.57	1.29	0.01	2.99	0.28	0	0.29	0.59	1.82	16.2	1.98	6.49	6.88
Refseq3FTRNA	tryptic1	20ppm	No	XCorr	No	cpu	118.69	0	5.58	1.3	0	44.12	5.76	0	3.95	8.44	25.97	66.25	1.48	58.3	6.48
Refseq3FTRNA	tryptic1	20ppm	No	PeakMatch	No	gpu	32	0.36	5.57	0.47	0	3.01	0.28	0	0.34	0.59	1.79	21.56	1.98	11.98	6.74
Refseq3FTRNA	tryptic1	20ppm	No	PeakMatch	No	cpu	886.63	0	5.51	0.45	0	43.07	5.71	0	3.81	8.02	25.53	836.21	1.48	828.41	6.32
Refseq3FTRNA	tryptic1	20ppm	No	PeakBackground	No	gpu	31.34	0.5	5.79	0.52	0	2.92	0.22	0	0.31	0.66	1.73	20.58	1.99	10.89	6.87
Refseq3FTRNA	tryptic1	20ppm	No	PeakBackground	No	cpu	1495.64	0	5.52	0.47	0	42.93	5.71	0	3.88	7.88	25.46	1445.33	1.42	1437.5	6.41
Refseq3FTRNA	tryptic1	1Da	No	XCorr	No	gpu	164.72	0.34	5.46	1.28	0	17.17	0.4	0	0.48	1.01	15.28	139.28	18.55	56.21	59.29
Refseq3FTRNA	tryptic1	1Da	No	XCorr	No	cpu	824.79	0	5.55	1.3	0	229.87	5.72	0	4.04	7.78	212.32	583.61	12.59	515.47	55.53
Refseq	tryptic1	20ppm	Yes	XCorr	No	gpu	6.89	0.33	1.4	1.3	0.01	0.37	0.04	0	0.04	0.1	0.19	2.59	0.29	1.05	1.12
Refseq	tryptic1	20ppm	Yes	XCorr	No	cpu	23.84	0	1.41	1.3	0.01	6.39	0.42	0	0.64	1.48	3.85	13.79	0.22	12.49	1.07
Refseq	tryptic1	20ppm	Yes	PeakMatch	No	gpu	6.92	0.39	1.46	0.48	0.01	0.35	0.02	0	0.04	0.09	0.2	3.33	0.28	1.83	1.1
Refseq	tryptic1	20ppm	Yes	PeakMatch	No	cpu	137.3	0	1.42	0.46	0.01	6.36	0.41	0	0.67	1.45	3.83	128.13	0.22	126.85	1.05
Refseq	tryptic1	20ppm	Yes	PeakBackground	No	gpu	6.65	0.35	1.38	0.51	0.01	0.35	0.03	0	0.03	0.1	0.19	3.18	0.28	1.68	1.1
Refseq	tryptic1	20ppm	Yes	PeakBackground	No	cpu	227.43	0	1.44	0.49	0.01	6.34	0.41	0	0.65	1.46	3.81	218.24	0.22	216.95	1.07
Refseq	tryptic1	1Da	Yes	XCorr	No	gpu	46.53	0.32	1.38	1.28	0.01	2.73	0.1	0	0.05	0.23	2.35	39.86	4.35	16.73	17.36
Refseq	tryptic1	1Da	Yes	XCorr	No	cpu	292.25	0	1.44	1.29	0.02	46.6	0.38	0	0.67	0.7	44.83	241.04	3.73	220.7	16.6
Refseq	tryptic1	1Da	Yes	PeakMatch	No	gpu	60.17	0.34	1.39	0.47	0	2.75	0.05	0	0.05	0.18	2.48	54.32	4.32	31.42	17.05
Refseq	tryptic1	1Da	Yes	PeakMatch	No	cpu	2324.94	0	1.39	0.45	0.02	45.89	0.34	0.02	0.62	0.62	44.26	2275.36	3.45	2255.73	16.13
Refseq	tryptic1	1Da	Yes	PeakBackground	No	gpu	61.33	0.38	1.42	0.5	0	4.08	0.7	0	0.28	0.24	2.86	53.97	5.13	29.45	17.88
Refseq	tryptic1	1Da	Yes	PeakBackground	No	cpu	3927.82	0	1.39	0.48	0	45.8	0.32	0	0.69	0.62	44.16	3878.33	3.42	3858.57	16.34
Refseq	semi	20ppm	Yes	XCorr	No	gpu	79.59	0.34	1.39	1.3	0.14	9.08	0.45	0	0.87	2.79	4.97	66.32	7.69	28.73	26.27
Refseq	semi	20ppm	Yes	XCorr	No	cpu	528.26	0	1.39	1.3	0.14	151.23	4.74	0	17.55	39.5	89.42	371.48	6.2	340.92	24.37
Refseq	semi	20ppm	Yes	PeakMatch	No	gpu	99.53	0.37	1.37	0.48	0.14	8.91	0.25	0	0.76	2.93	4.96	87.31	7.72	50.39	25.66
Refseq	semi	20ppm	Yes	PeakMatch	No	cpu	3586.59	0	1.37	0.47	0.13	147.87	4.73	0	17.26	37.74	88.15	3434.1	5.96	3404.32	23.8
Refseq	semi	20ppm	Yes	PeakBackground	No	gpu	95.16	0.36	1.39	0.5	0.14	8.83	0.38	0	0.76	2.8	4.89	82.98	7.54	46.26	25.69
Refseq	semi	20ppm	Yes	PeakBackground	No	cpu	5967.92	0	1.37	0.5	0.13	148.22	4.67	0	17.22	37.96	88.37	5815.04	6	5784.77	24.26
Refseq	nonspec	20ppm	Yes	XCorr	No	gpu	384.18	0.31	1.48	1.28	1	45.23	2.46	0	3.87	15.89	23	333.4	37.36	156.37	121.93
Refseq	nonspec	20ppm	Yes	XCorr	No	cpu	2489.73	0	1.4	1.3	1.01	717.36	10.9	0	89.59	205.64	411.11	1758.75	29.09	1616.46	113.15
Refseq3FTRNA	tryptic1	20ppm	Yes	XCorr	No	gpu	48.38	0.32	5.57	1.3	0.01	5.04	0.26	0	0.49	1.63	2.66	35.04	4.05	14.72	14.29
Refseq3FTRNA	tryptic1	20ppm	Yes	XCorr	No	cpu	300.08	0	5.52	1.26	0.02	92.09	5.7	0	10.15	22.62	53.62	199.38	3.22	180.96	15.19
Refseq3FTRNA	tryptic1	20ppm	Yes	PeakMatch	No	gpu	60.17	0.34	5.55	0.47	0	5.02	0.34	0	0.38	1.61	2.7	47.75	4.06	27.26	14.46
Refseq3FTRNA	tryptic1	20ppm	Yes	PeakMatch	No	cpu	1972.19	0	5.52	0.44	0	91.79	5.68	0	10.23	22.61	53.27	1872.63	3.15	1856.16	13.32
Refseq3FTRNA	tryptic1	20ppm	Yes	PeakBackground	No	gpu	58.2	0.34	5.55	0.5	0	5.1	0.34	0	0.39	1.59	2.78	45.65	4.04	25.02	14.64
Refseq3FTRNA	tryptic1	20ppm	Yes	PeakBackground	No	cpu	3289.55	0	5.52	0.47	0.02	91.51	5.6	0	10.26	22.48	53.16	3190.12	3.14	3173.58	13.4

Appendix C

SUPPLEMENTARY MATERIALS FOR COMPREHENSIVE PROFILING OF MODIFIED PEPTIDES

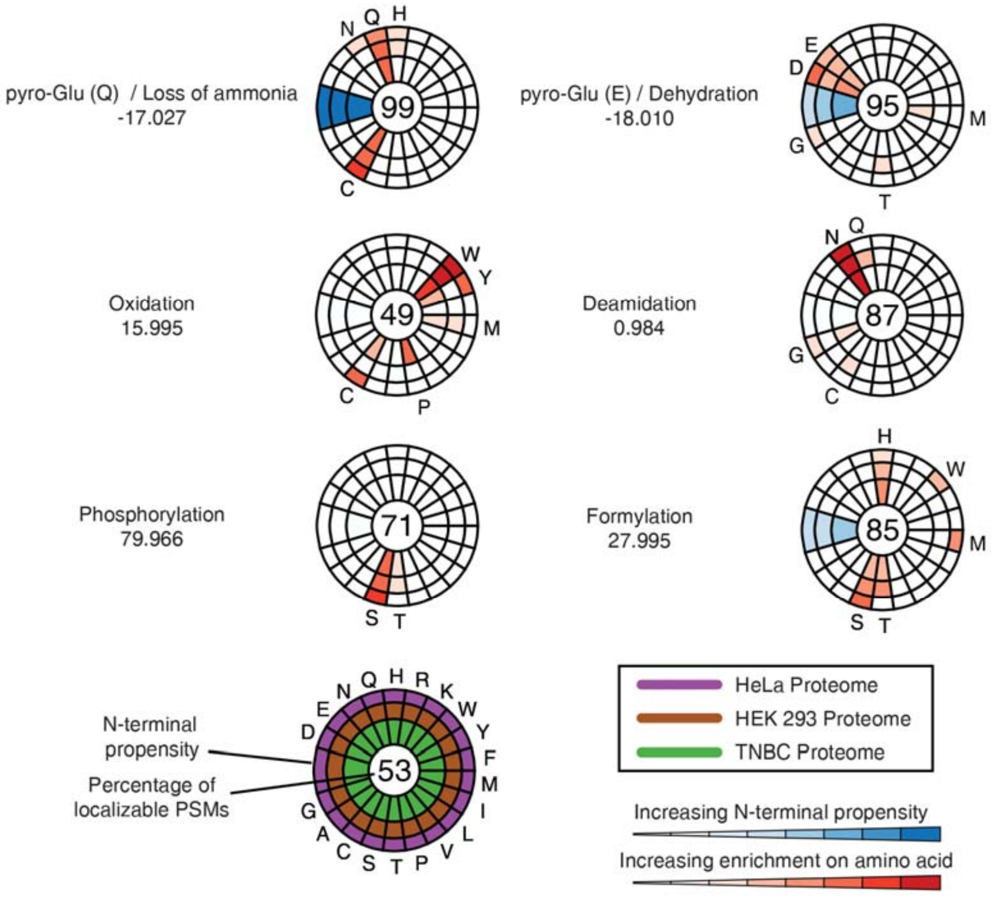


Figure C-1 Localization profiles are consistent across experiments.
Common modifications were selected and amino acid localization enrichment was calculated separately for each dataset. Amino acid localizations were largely consistent across each dataset despite the differences in modification rates.

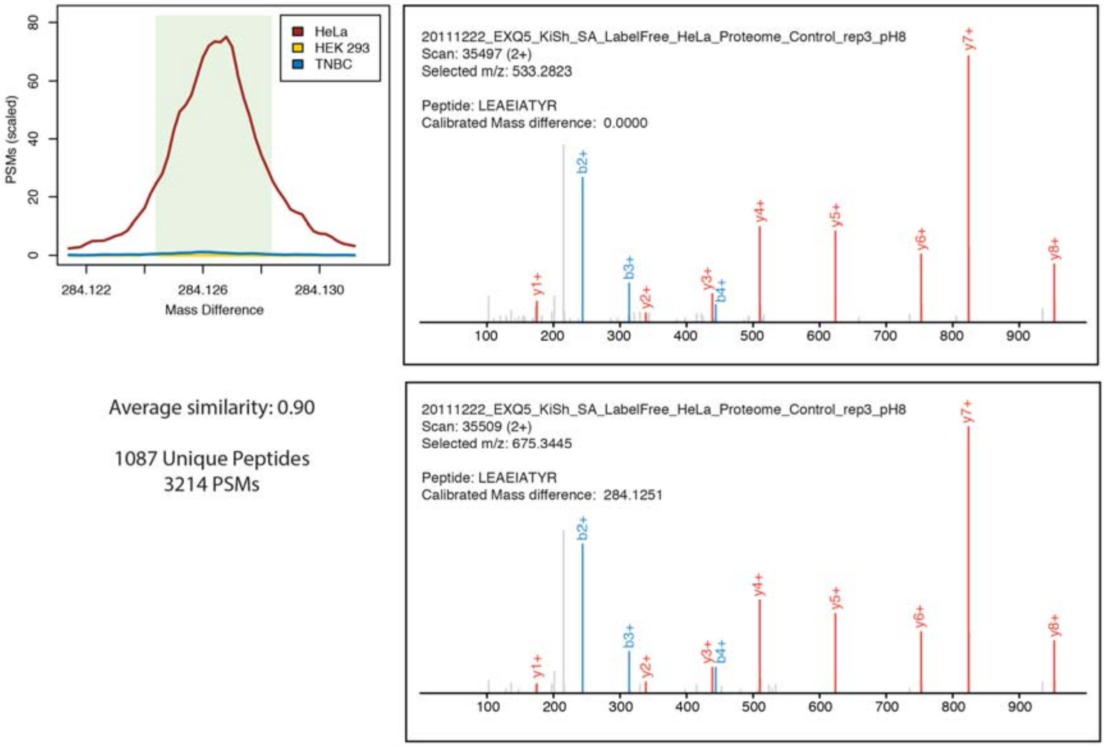


Figure C-2 Highly similar spectra pair for peptide LEAEIATYR with precursor mass difference of 284.126.
3214 PSMs (corresponding to 1087 unique peptides) were identified in the mass difference bin of 284.126 Da. These PSMs were predominantly observed in the HeLa dataset and were shown to have a spectral similarity score of 0.90 (indicating that the spectra of mass shifted peptides are highly similar to that of corresponding unmodified peptides). Here, we selected a pair of PSMs that were both identified to be the peptide LEAEIATYR in the same LC-MS/MS run. Despite their highly similar fragmentation patterns and few unmatched fragments, they were observed with precursor masses that differ by 284.1251 Da. The full y-ion series was successfully matched, which when overlapped with the matched b-2, b-3, and b-4 ions, rules out the possibility of a modified residue in the fragmentation spectrum.

Table C-1 Mass shift localization by dataset.

BinStart	Lower bound of mass shift in bin
BinEnd	Upper bound of mass shift in bin
LocPSMs	Number of PSMs in which the mass shift could be localized
TotalPSMs	Total number of PSMs within this bin
Nterm	Rate at which mass shift is localized on N-terminal end of peptide (percentage)
TopAA1	1st highest enriched amino acid with enrichment ratio
TopAA2	2nd highest enriched amino acid with enrichment ratio
TopAA3	3rd highest enriched amino acid with enrichment ratio

		HEK293						TNBC						HeLa					
BinStart	BinEnd	LocPSMs	TotalPSMs	Nterm	TopAA1	TopAA2	TopAA3	LocPSMs	TotalPSMs	Nterm	TopAA1	TopAA2	TopAA3	LocPSMs	TotalPSMs	Nterm	TopAA1	TopAA2	TopAA3
-0.0022	0.0018	2131	300606	0.38	R - 1.31	K - 1.30	H - 1.30	32626	3301510	0.5	K - 1.32	D - 1.18	H - 1.16	9360	912117	0.44	D - 1.76	E - 1.58	M - 1.33
1.0006	1.0046	6556	27106	0.67	P - 1.69	D - 1.37	E - 1.34	133981	390685	0.51	D - 1.52	E - 1.42	P - 1.23	603	2895	0.86	Y - 1.91	D - 1.87	E - 1.66
43.0038	43.0078	8089	11996	61.15	H - 1.78	A - 1.68	V - 1.67	46010	69706	60.41	I - 2.29	V - 2.12	L - 1.73	7	15	40	A - 4.09	E - 4.03	K - 2.51
0.982	0.986	3406	3891	11.67	N - 16.44	Q - 2.72	C - 2.21	34225	40417	11.9	N - 15.77	G - 1.80	C - 1.16	13070	13897	12.84	N - 19.96	G - 1.51	D - 0.61
-17.0288	-17.0248	4699	4752	69.3	Q - 6.20	C - 5.55	H - 1.57	31329	31840	73.88	C - 6.47	Q - 5.82	N - 1.20	6238	6312	73.13	C - 7.98	Q - 4.34	H - 1.90
52.9108	52.9148	0	1	0	A - 0.00	A - 0.00	A - 0.00	1103	4495	0.56	D - 2.48	E - 2.40	N - 1.79	16454	24269	0.41	D - 2.94	E - 2.23	Y - 1.49
-128.097	-128.093	740	1072	1.49	K - 2.52	F - 2.42	E - 2.40	14733	21435	3	F - 2.39	I - 2.05	Y - 1.78	5114	7177	2.22	E - 3.91	F - 2.85	Y - 2.19
2.0032	2.0072	798	4889	0.14	P - 2.01	Q - 1.49	D - 1.39	37025	126787	0.18	D - 1.56	P - 1.43	E - 1.28	43	182	0.55	Y - 3.00	M - 2.57	W - 2.49
128.0928	128.0968	1174	1485	40.81	K - 2.38	R - 1.78	I - 1.46	33796	40120	49.05	K - 1.81	R - 1.73	I - 1.61	2479	2711	55.85	Y - 2.47	D - 1.90	I - 1.89
15.9928	15.9968	744	1764	6.58	W - 50.37	M - 1.93	Y - 1.38	2977	6954	7.71	W - 8.84	P - 5.44	Y - 2.37	3912	6828	11.15	W - 28.95	Y - 5.23	C - 5.13
31.988	31.992	1495	1794	43.87	W - 65.51	Q - 0.34	I - 0.32	1124	2280	5.57	W - 41.41	P - 2.05	Y - 1.76	2027	4176	5.56	W - 59.58	D - 1.01	Y - 0.63
301.9844	301.9884	54	3417	0	R - 6.11	H - 2.08	D - 1.83	12	943	0	P - 8.68	L - 1.78	R - 1.63	0	2	0	A - 0.00	A - 0.00	A - 0.00
27.993	27.997	2352	2712	28.54	S - 4.57	T - 3.42	H - 2.37	7579	8868	42.68	H - 4.01	S - 3.26	T - 2.37	295	477	27.67	S - 5.99	M - 4.56	W - 2.59
79.9644	79.9684	860	1216	3.37	S - 7.24	T - 2.11	P - 1.18	4326	6588	5.89	S - 6.91	T - 1.88	P - 1.40	3001	3743	3.82	S - 8.00	T - 1.19	D - 1.17
53.9166	53.9206	1734	2834	1.52	D - 2.97	N - 2.27	E - 1.59	272	976	0.41	D - 2.44	E - 1.89	N - 1.84	260	291	0	C - 3.62	A - 2.80	D - 2.41
183.0332	183.0372	1870	2177	15.3	Y - 20.47	K - 4.42	G - 0.44	6036	6988	15.93	Y - 25.05	K - 1.97	G - 0.50	67	73	8.22	Y - 27.43	K - 1.92	E - 0.84
57.0194	57.0234	403	821	39.83	I - 2.79	H - 2.71	M - 1.63	293	638	25.39	P - 3.42	M - 2.10	H - 2.00	2348	5415	10.38	H - 9.83	K - 2.94	M - 2.64
23.956	23.96	0	0	0	A - 0.00	A - 0.00	A - 0.00	50	83	0	V - 3.58	D - 3.03	I - 1.93	5402	6167	0.28	D - 3.55	E - 2.56	Y - 1.78
-18.0124	-18.0084	417	462	44.59	E - 3.31	D - 2.77	T - 1.86	1802	1908	54.66	D - 3.40	E - 3.04	M - 1.72	3167	3318	36.83	D - 6.61	E - 2.38	G - 1.86
156.0992	156.1032	529	581	53.7	V - 1.73	R - 1.55	D - 1.46	10264	10360	79.75	I - 1.60	A - 1.54	D - 1.45	1172	1179	75.23	Q - 2.47	D - 2.18	Y - 1.79
284.1248	284.1288	0	0	0	A - 0.00	A - 0.00	A - 0.00	4	292	0.34	M - 5.73	N - 3.53	S - 2.97	5	3030	0.1	I - 5.85	T - 2.43	S - 2.37
162.1238	162.1278	0	0	0	A - 0.00	A - 0.00	A - 0.00	0	3	0	A - 0.00	A - 0.00	A - 0.00	24	2900	0.66	F - 5.84	W - 3.35	Y - 3.19
17.0238	17.0278	8	222	0	N - 8.15	I - 2.92	A - 2.21	124	4271	0.14	W - 2.05	M - 2.01	T - 1.92	21	2479	0	D - 3.50	G - 2.54	Q - 2.36
234.0722	234.0762	0	0	0	A - 0.00	A - 0.00	A - 0.00	0	4	0	A - 0.00	A - 0.00	A - 0.00	6	2717	0.15	Y - 22.34	I - 3.90	E - 3.52
37.9452	37.9492	150	194	0.52	D - 3.54	V - 1.78	P - 1.49	3171	3449	1.8	D - 3.34	E - 1.94	V - 1.52	2836	3056	0.59	D - 2.48	E - 2.31	N - 1.68
306.0932	306.0972	1	11	0	V - 16.68	A - 0.00	A - 0.00	1	6	0	K - 8.79	A - 7.15	A - 0.00	10	2853	0.25	L - 7.08	Y - 0.96	V - 0.83
44.0064	44.0104	251	972	15.53	V - 1.83	I - 1.42	K - 1.36	3107	8473	21.74	V - 1.74	I - 1.63	Y - 1.45	0	2	0	A - 0.00	A - 0.00	A - 0.00
-15.9976	-15.9936	52	241	12.45	M - 13.96	Y - 7.73	N - 3.55	2210	6847	9.36	M - 27.04	N - 0.85	S - 0.74	350	907	13.23	M - 30.99	N - 1.88	D - 0.91
302.9876	302.9916	24	994	0.1	C - 2.25	P - 1.81	V - 1.53	2	130	0.77	C - 8.94	W - 8.03	G - 5.69	0	0	0	A - 0.00	A - 0.00	A - 0.00
-71.0388	-71.0348	261	273	54.58	A - 8.63	V - 0.91	E - 0.91	2319	2427	60.36	A - 8.08	D - 3.02	W - 0.82	736	744	52.28	A - 8.82	D - 4.89	Y - 0.84
249.979	249.983	0	732	0	A - 0.00	A - 0.00	A - 0.00	0	21	0	A - 0.00	A - 0.00	A - 0.00	0	1	0	A - 0.00	A - 0.00	A - 0.00
-113.086	-113.082	337	362	48.07	I - 5.34	L - 4.36	F - 1.07	1772	1890	46.88	L - 4.88	I - 4.67	T - 0.67	303	321	58.26	I - 4.43	L - 3.88	F - 1.73
21.9784	21.9824	273	288	0.35	D - 2.38	N - 1.86	A - 1.75	7718	7814	0.47	D - 3.55	E - 2.00	A - 1.46	193	208	0.96	D - 3.58	E - 2.94	I - 1.82
241.1768	241.1808	29	140	20	D - 6.66	G - 3.93	F - 1.93	2505	5150	38.95	D - 4.44	E - 2.64	G - 2.31	59	306	18.3	D - 4.98	G - 4.80	M - 3.63
-145.124	-145.12	270	270	0.37	C - 10.12	Q - 9.95	Y - 2.01	1655	1683	1.07	C - 17.68	Q - 7.13	Y - 1.23	577	582	0.34	C - 16.50	Q - 5.76	Y - 1.82
14.0136	14.0176	192	236	15.68	H - 5.72	V - 4.65	I - 2.50	1463	3001	14.33	V - 4.88	D - 2.89	H - 2.50	300	490	20.61	V - 8.67	H - 2.54	D - 1.57
1.9862	1.9902	508	947	1.58	N - 13.97	Q - 1.89	G - 1.39	7936	13806	2.51	N - 10.81	G - 1.60	Q - 1.04	36	126	0.79	N - 14.17	M - 2.55	G - 1.00
-2.018	-2.014	245	352	51.14	M - 15.89	C - 5.94	W - 4.43	1945	3148	40.28	W - 18.12	M - 13.08	C - 1.54	331	415	27.71	W - 38.76	M - 4.49	Y - 3.26
151.9936	151.9976	10	14	7.14	C - 44.71	A - 0.00	A - 0.00	8310	8793	16.32	C - 37.44	G - 0.34	D - 0.25	7	7	28.57	C - 31.94	G - 4.34	A - 0.00
3.0056	3.0096	106	634	0.16	P - 2.53	E - 1.46	Q - 1.39	5659	20899	0.18	P - 1.49	Q - 1.35	C - 1.32	4	20	0	I - 7.80	V - 3.13	N - 1.77
-85.0914	-85.0874	323	344	10.17	I - 3.62	E - 2.20	L - 1.89	1563	1710	1.87	I - 4.20	L - 2.98	E - 1.74	0	0	0	A - 0.00	A - 0.00	A - 0.00
-16.0266	-16.0226	374	456	48.9	Q - 5.27	C - 4.37	N - 2.01	4166	4819	61.69	Q - 5.25	C - 4.73	H - 1.55	28	36	52.78	C - 9.95	H - 2.83	Q - 2.16
16.9948	16.9988	187	555	1.26	W - 26.88	M - 7.28	C - 1.25	815	2093	3.49	W - 6.54	M - 6.21	P - 2.36	14	35	0	W - 19.13	M - 14.41	Y - 6.02
-1.0056	-1.0016	12	534	0.19	I - 4.78	V - 2.55	Q - 2.38	201	3791	1.98	M - 4.83	W - 3.01	V - 1.48	12	695	0.43	Q - 6.82	W - 6.70	N - 1.86
54.9198	54.9238	176	546	0.73	D - 2.46	N - 2.20	E - 1.75	57	363	0.28	D - 2.12	A - 1.88	M - 1.87	13	18	0	I - 5.10	E - 3.20	C - 2.15
80.9682	80.9722	215	436	0.69	S - 3.85	P - 2.34	D - 1.54	1567	2952	1.8	S - 4.39	P - 2.14	T - 1.71	2	5	0	S - 7.12	N - 2.83	I - 2.34
204.134	204.138	0	0	0	A - 0.00	A - 0.00	A - 0.00	0	4	0	A - 0.00	A - 0.00	A - 0.00	3	1001	0.2	F - 18.69	H - 6.39	V - 1.39
131.0376	131.0416	1	16	0	S - 11.86	A - 0.00	A - 0.00	88	2072	2.9	K - 7.85	I - 3.46	R - 1.26	16	566	1.77	W - 10.04	K - 5.71	G - 1.90
125.8948	125.8988	331	349	21.2	Y - 31.25	W - 4.16	G - 0.42	119	128	28.13	Y - 31.37	E - 1.26	N - 0.83	2	2	100	Y - 38.29	A - 0.00	A - 0.00
184.035	184.039	244	404	6.93	Y - 17.91	K - 3.42	G - 0.72	1114	1496	6.15	Y - 20.02	K - 1.59	G - 0.74	0	2	0	A - 0.00	A - 0.00	A - 0.00
129.0954	129.0994	121	160	24.38	E - 2.14	I - 1.84	V - 1.74	4575	5650	34.48	D - 1.55	E - 1.46	V - 1.35	2	3	0	Q - 5.22	V - 4.17	I - 2.92
216.0982	216.1022	0	1	0	A - 0.00	A - 0.00	A - 0.00	4	11	9.09	F - 4.09	P - 3.94	V - 3.82	17	896	1.34	K - 4.48	G - 2.78	A - 2.24
-0.9856	-0.9816	20	533	0.94	D - 3.94	P - 2.39	E - 2.04	553	3686	2.22	D - 8.22	E - 1.93	Y - 1.28	152	335	31.04	P - 5.19	D - 2.86	M - 1.78
197.0434	197.0474	87	115	0	E - 12.72	A - 0.82	S - 0.48	823	953	0.21	E - 11.60	S - 1.12	A - 0.99	352	359	0.28	E - 12.13	S - 1.35	A - 0.31
3.9926	3.9966	166	239	5.86	W - 68.29	N - 0.65	I - 0.42	390	1111	0.72	W - 8.93	N - 4.30	P - 1.59	236	278	7.19	W - 67.15	P - 0.47	I - 0.43
203.078	203.082	4	183	0	H - 19.16	T - 3.03	S - 2.97	58	999	1.5	N - 5.16	P - 3.43	T - 3.10	11	253	0	P - 3.39	S - 3.31	D - 3.18
-14.0176	-14.0136	79	117	27.35	I - 8.81	E - 3.07	Y - 2.62	829	1336	19.24	I - 6.62	T - 3.48	E - 1.89	263	376	3.19	I - 10.74	E - 2.58	A - 1.82
-1.034	-1.03	144	455	26.15	M - 11.73	C - 5.74	Q - 3.30	948	2781	27.58	M - 2.84	C - 2.46	L - 1.84	240	407	50.86	M - 3.90	C - 3.38	L - 2.72
-99.071	-99.067	202	216	56.94	V - 9.88	C - 1.44	Y - 1.12	790	832	63.58	V - 8.09	I - 2.59	N - 1.60	35	39	28.21	V - 11.96	W - 3.06	K - 0.75
41.0244	41.0284	9	77	7.79	P - 12.25	T - 2.02	D - 1.17	44	138	25.36	P - 6.65	S - 4.58	V - 1.33	92	654	7.49	P - 5.40	E - 2.35	C - 2.19
173.0486	173.0526	0	46	0	A - 0.00	A - 0.00	A - 0.00	176	1939	8.25	E - 7.27	N - 3.85	D - 3.71	31	308	10.06	M - 28.10	N - 5.47	E - 1.82
0.9462	0.9502	7	342	0.88	K - 10.04	V - 5.96	G - 1.08	113	1793	1.12	K - 5.65	Y - 3.59	V - 1.90	30	706	0.71	K - 7.91	L - 3.28	D - 1.81
28.0294	28.0334	166	185	13.51	R - 6.26	G - 2.24	P - 1.84	1180	1385	15.09	R - 4.46	G - 2.75	A - 1.93	72	78	19.23	K - 7.81	R - 1.69	S - 1.39
248.1242	248.1282	0	4	0	A - 0.00	A - 0.00	A - 0.00	0	8	0	A - 0.00	A - 0.00	A - 0.00	3	695	0.14	I - 7.80	A - 4.77	L - 3.37
28.9954	28.9994	239	361	5.54	S - 4.14	T - 2.94	D - 1.10	1346	1832	13.43	S - 2.92	T - 2.50	D - 1.04	4	9	0	I - 10.24	S - 2.72	C - 1.86
-328.213	-328.209	20	24	29.17	K - 6.22	T - 2.73	S - 2.13	275	389	23.14	K - 9.89	I - 2.47	T - 1.81	478	527	57.5	K - 5.89	I - 5.61	S - 3.04
-156.103	-156.099	25	41	7.32	R - 5.53	C - 5.37	W - 3.21	829	1352	10.72	R - 9.34	C - 2.15	Y - 1.50	316	336	1.79	R - 14.19	A - 1.32	C - 0.92
-229.145	-229.141	32	34	41.18	K - 6.50	T - 4.93	P - 4.02	587	694	67.29	T - 8.02	K - 6.92	P - 2.18	438	457	34.14	K - 9.38	T - 4.51	P - 2.42
-127.113	-127.109	66	75	1.33	N - 15.46	C - 5.65	G - 2.45	641	749	1.34	N - 18.98	C - 3.44	G - 1.86	382	436	0	N - 17.52	G - 1.41	M - 1.33

11.9976	12.0016	79	86	73.26	H - 14.11	M - 5.24	N - 3.16	751	869	71.69	W - 28.62	H - 5.82	M - 3.09	219	270	76.67	W - 51.66	H - 5.06	Y - 0.82
43.988	43.992	172	221	26.24	W - 42.76	N - 2.36	M - 1.15	563	1064	7.89	W - 20.89	N - 4.46	A - 1.96	0	0	0	A - 0.00	A - 0.00	A - 0.00
-170.108	-170.104	152	161	63.35	V - 3.24	I - 2.55	A - 2.26	298	366	50.55	G - 2.36	V - 2.07	P - 1.95	68	83	54.22	V - 2.71	F - 2.20	M - 1.99
284.1944	284.1984	98	99	88.89	F - 6.12	G - 4.51	E - 1.74	1003	1027	83.84	F - 3.46	T - 2.39	E - 2.36	162	162	72.84	F - 9.05	G - 5.07	A - 2.15
-257.14	-257.136	31	45	51.11	K - 6.05	E - 5.30	P - 2.12	328	554	45.85	K - 5.47	E - 5.12	P - 1.38	324	396	48.48	K - 6.20	E - 4.55	S - 2.50
32.9904	32.9944	134	295	6.78	W - 54.24	D - 0.90	C - 0.73	254	550	5.64	W - 21.84	P - 2.98	G - 1.46	10	28	3.57	W - 68.29	V - 1.67	S - 0.59
-57.0236	-57.0196	47	51	52.94	H - 8.02	G - 6.51	V - 2.22	1938	2045	22.1	C - 25.50	G - 2.33	H - 0.99	145	157	44.59	G - 7.35	H - 4.31	F - 2.43
-0.0382	-0.0342	11	291	2.06	M - 6.95	V - 3.79	I - 2.66	154	2314	1.64	Y - 1.97	K - 1.93	V - 1.84	8	580	0.69	Y - 7.18	T - 2.65	F - 2.34
171.099	171.103	39	123	24.39	V - 3.21	K - 1.92	I - 1.80	525	1373	28.55	V - 4.38	I - 1.82	L - 1.39	0	3	0	A - 0.00	A - 0.00	A - 0.00
-199.134	-199.13	21	25	36	K - 4.88	A - 4.88	P - 3.13	431	552	28.99	A - 6.00	K - 4.54	P - 2.66	337	372	10.48	A - 9.53	K - 2.43	M - 1.49
-129.044	-129.04	15	29	10.34	E - 11.12	H - 2.56	D - 1.40	185	297	20.54	E - 10.25	M - 1.36	F - 1.07	376	399	63.66	E - 8.27	Y - 3.45	F - 1.09
0.0346	0.0386	10	292	1.71	N - 6.36	I - 2.92	D - 2.52	2262	5958	0.97	E - 1.50	F - 1.43	Q - 1.41	5	340	1.47	Q - 12.54	F - 5.61	N - 1.41
378.1126	378.1166	0	3	0	A - 0.00	A - 0.00	A - 0.00	1	4	25	N - 8.07	F - 4.01	L - 2.89	5	691	0	I - 14.04	P - 3.15	G - 0.91
-9.039	-9.035	117	135	11.85	C - 38.41	E - 0.66	Q - 0.51	984	1242	7.81	C - 38.57	E - 0.55	D - 0.24	27	39	7.69	C - 38.64	N - 1.57	E - 0.78
-87.034	-87.03	49	50	38	S - 8.84	H - 0.78	L - 0.60	529	578	45.67	S - 7.67	D - 1.22	V - 0.98	238	256	46.09	S - 7.95	D - 2.19	V - 1.37
-184.123	-184.119	89	94	48.94	A - 4.36	L - 2.93	I - 2.26	209	234	60.68	A - 4.83	L - 2.77	P - 1.96	186	191	68.59	A - 4.45	Y - 3.77	L - 2.98
-200.118	-200.114	102	106	49.06	V - 2.88	T - 2.39	P - 2.32	271	346	47.4	T - 3.22	L - 2.54	S - 2.19	107	124	33.06	V - 3.63	L - 2.73	T - 2.40
170.0928	170.0968	0	1	0	A - 0.00	A - 0.00	A - 0.00	4	66	3.03	F - 7.01	N - 3.53	I - 2.92	0	467	0	A - 0.00	A - 0.00	A - 0.00
244.129	244.133	3	5	40	F - 7.79	L - 2.81	D - 2.33	85	131	51.15	D - 2.00	E - 1.84	G - 1.75	0	521	0	A - 0.00	A - 0.00	A - 0.00
113.0816	113.0856	20	83	4.82	R - 5.44	K - 3.51	M - 2.58	471	1001	3.3	K - 5.20	R - 4.82	N - 1.21	5	92	4.35	M - 23.69	K - 3.51	D - 3.50
-114.044	-114.04	57	61	68.85	N - 13.38	F - 4.67	I - 3.63	356	390	71.28	N - 11.32	F - 2.79	G - 2.50	247	257	89.88	N - 13.40	F - 3.03	I - 1.37
260.125	260.129	0	1	0	A - 0.00	A - 0.00	A - 0.00	5	21	19.05	V - 3.64	D - 3.21	E - 2.14	1	470	0	E - 14.09	A - 0.00	A - 0.00
68.9058	68.9098	0	0	0	A - 0.00	A - 0.00	A - 0.00	3	29	3.45	M - 3.50	Y - 2.92	G - 2.49	291	504	7.54	M - 15.80	D - 1.59	N - 1.50
157.1022	157.1062	93	103	41.75	D - 2.49	Q - 1.86	G - 1.52	2652	2693	75.01	F - 1.74	D - 1.56	V - 1.51	0	0	0	A - 0.00	A - 0.00	A - 0.00
269.1834	269.1874	35	36	80.56	D - 4.90	E - 2.01	G - 1.73	1185	1199	82.9	D - 3.82	E - 2.43	G - 1.83	162	162	92.59	E - 2.92	C - 2.32	H - 2.03
-112.102	-112.098	15	34	0	M - 22.41	W - 5.36	Q - 1.86	131	245	2.45	M - 16.63	E - 2.07	Q - 2.02	205	288	0.69	W - 15.42	M - 13.88	Q - 2.86
-128.061	-128.057	61	67	40.3	Q - 7.14	F - 2.34	A - 1.99	298	372	29.57	Q - 11.17	H - 1.32	E - 1.29	133	154	27.27	A - 6.46	Q - 5.75	H - 1.91
42.0084	42.0124	27	70	24.29	M - 10.75	T - 2.76	K - 2.60	419	989	33.27	M - 11.87	A - 3.99	W - 2.36	92	154	40.26	A - 5.23	K - 4.68	M - 1.99
-215.129	-215.125	19	21	57.14	S - 6.76	K - 6.32	P - 0.83	192	365	12.33	S - 5.41	K - 5.01	Y - 1.13	265	297	13.13	S - 7.16	K - 2.61	W - 1.01
-115.029	-115.025	60	63	68.25	D - 11.80	P - 2.98	W - 1.12	1077	1135	75.51	D - 11.67	P - 3.33	E - 0.53	85	89	43.82	D - 15.06	G - 1.52	Q - 0.74
49.9982	50.0022	1	2	0	I - 5.85	R - 4.41	P - 3.94	3	43	0	Y - 4.25	M - 2.48	P - 2.28	22	356	0.56	Y - 20.01	F - 11.26	Q - 0.95
0.0454	0.0494	5	304	0.33	M - 9.17	V - 3.34	N - 2.83	866	3147	0.32	D - 1.48	E - 1.44	I - 1.41	2	308	0.32	L - 5.06	F - 3.51	I - 2.92
0.9356	0.9396	2	335	0	I - 5.85	Q - 5.22	E - 3.52	26	1546	0.19	V - 3.35	I - 2.56	Q - 1.77	4	546	0	N - 7.06	T - 6.83	V - 4.17
210.1596	210.1636	0	0	0	A - 0.00	A - 0.00	A - 0.00	2714	2769	97.36	I - 3.03	L - 2.60	F - 2.21	1	2	50	I - 23.40	A - 0.00	A - 0.00
53.8952	53.8992	0	1	0	A - 0.00	A - 0.00	A - 0.00	40	215	0.47	V - 4.36	I - 3.25	D - 2.37	269	392	1.53	N - 5.27	V - 4.46	D - 1.91
308.0804	308.0844	0	9	0	A - 0.00	A - 0.00	A - 0.00	3	90	2.22	Q - 7.96	H - 6.39	V - 3.58	0	352	0	A - 0.00	A - 0.00	A - 0.00
13.977	13.981	28	49	32.65	W - 24.96	P - 1.69	I - 1.25	417	527	8.54	W - 29.22	P - 5.60	Y - 4.85	154	292	0.68	W - 5.65	P - 3.25	I - 2.68
-241.182	-241.178	24	31	16.13	I - 5.36	L - 4.85	K - 2.81	277	446	32.74	K - 6.23	I - 5.14	L - 3.26	206	245	46.12	I - 8.31	K - 5.47	L - 2.46
420.0486	420.0526	0	2	0	A - 0.00	A - 0.00	A - 0.00	0	2703	0	A - 0.00	A - 0.00	A - 0.00	0	2	0	A - 0.00	A - 0.00	A - 0.00
-163.065	-163.061	29	29	89.66	Y - 20.46	L - 2.27	M - 1.58	308	325	76	Y - 25.81	M - 2.60	H - 0.84	218	234	80.77	Y - 21.62	L - 2.17	A - 1.27
-286.166	-286.162	30	31	54.84	K - 5.27	T - 4.45	A - 3.02	262	312	41.67	T - 6.74	K - 4.50	A - 2.39	202	220	49.09	K - 4.72	A - 4.03	S - 2.79
2.9872	2.9912	98	255	0.78	N - 9.22	G - 2.13	P - 1.21	2621	5135	1.46	N - 8.79	G - 1.48	Q - 1.05	6	16	0	N - 4.71	T - 4.55	C - 3.73
215.125	215.129	8	29	10.34	D - 4.26	S - 2.97	R - 2.57	1188	1599	58.97	D - 3.37	Q - 1.99	I - 1.86	68	78	82.05	D - 11.33	F - 6.47	Q - 2.20
-372.222	-372.218	11	11	0	K - 17.57	A - 0.00	A - 0.00	207	217	6.45	K - 15.03	F - 1.35	M - 1.14	257	259	0	K - 14.62	I - 2.32	M - 0.92
-146.107	-146.103	45	58	3.45	E - 6.26	M - 3.06	K - 1.93	108	134	0	E - 7.68	C - 1.66	M - 1.63	202	222	1.8	E - 7.90	W - 6.10	F - 1.59
-227.166	-227.162	34	41	56.1	V - 8.42	K - 4.74	P - 3.09	389	451	72.06	V - 7.89	K - 6.74	P - 1.71	136	148	41.89	V - 8.92	K - 5.52	N - 2.84
213.0982	213.1022	0	5	0	A - 0.00	A - 0.00	A - 0.00	1	79	1.27	F - 8.01	H - 5.48	N - 4.04	3	312	0	L - 3.93	N - 3.14	M - 2.55
-243.124	-243.12	9	18	33.33	D - 11.27	K - 3.58	A - 1.59	306	483	46.58	D - 10.76	K - 5.92	T - 0.61	137	184	70.65	D - 10.29	K - 7.97	P - 0.42
250.9802	250.9842	1	219	0	V - 5.56	D - 3.50	P - 2.63	1	16	6.25	V - 3.58	Q - 2.98	A - 2.04	0	1	0	A - 0.00	A - 0.00	A - 0.00
-212.155	-212.151	96	102	78.43	V - 6.57	I - 4.12	L - 2.74	144	146	67.81	I - 8.06	V - 4.88	L - 2.40	20	21	76.19	V - 6.46	I - 6.14	L - 3.41
-71.0752	-71.0712	43	57	0	I - 4.35	V - 4.27	F - 2.28	11	25	20	I - 7.23	S - 2.64	A - 2.43	79	188	0.53	F - 11.71	H - 7.47	M - 3.48
227.1618	227.1658	23	42	28.57	D - 9.09	E - 3.75	R - 1.48	710	1426	32.19	D - 4.01	R - 3.33	E - 2.52	9	25	28	M - 8.07	E - 5.61	N - 2.35
-131.043	-131.039	8	8	100	M - 25.79	A - 1.79	S - 1.48	221	350	47.43	M - 26.35	A - 1.63	C - 1.30	231	255	78.04	M - 26.98	S - 1.86	F - 1.83
16.9762	16.9802	45	74	32.43	M - 6.21	N - 4.79	G - 3.16	289	567	30.86	M - 4.60	G - 3.25	N - 3.25	104	164	39.63	N - 6.69	M - 5.04	G - 2.43
27.009	27.013	15	65	18.46	M - 8.25	P - 3.50	S - 1.98	316	666	27.18	M - 11.21	S - 2.49	H - 1.76	55	89	0	S - 10.29	W - 1.63	N - 1.52
-91.0116	-91.0076	5	5	0	C - 40.24	T - 1.82	A - 0.00	377	383	4.44	C - 38.13	H - 0.72	Y - 0.47	270	285	2.46	C - 40.17	Q - 0.35	D - 0.33
0.9242	0.9282	2	282	0	F - 14.02	I - 3.90	D - 3.50	34	1412	0.5	I - 3.77	Y - 3.38	C - 3.07	4	537	0.19	I - 3.90	L - 2.53	A - 2.50
-101.049	-101.045	16	21	61.9	T - 7.49	F - 2.34	I - 1.95	296	330	63.03	T - 8.98	L - 1.89	F - 1.85	189	209	22.97	T - 4.17	L - 3.68	F - 2.72
-142.077	-142.073	25	25	84	A - 8.58	M - 1.83	I - 1.25	65	83	53.01	A - 8.46	Y - 1.16	V - 1.16	214	217	81.11	A - 10.87	N - 2.20	Y - 2.15
-204.092	-204.088	38	42	50	F - 10.21	G - 7.72	M - 1.21	103	114	87.72	F - 10.33	G - 5.97	T - 2.01	137	148	44.59	F - 7.95	G - 5.52	A - 4.24
227.1362	227.1402	30	30	76.67	V - 7.60	E - 2.82	D - 2.10	736	743	92.46	V - 2.53	N - 2.38	A - 2.16	125	127	83.46	F - 7.23	I - 3.17	H - 2.91
-1.9814	-1.9774	6	154	0.65	M - 15.28	R - 3.24	P - 2.10	303	1758	5.29	M - 7.85	T - 1.54	G - 1.51	29	88	2.27	M - 7.02	I - 3.73	E - 1.42
-0.9494	-0.9454	7	281	0.36	V - 4.05	E - 3.62	N - 2.02	141	1582	1.58	E - 8.07	V - 1.15	G - 0.91	8	145	2.07	E - 7.63	T - 3.49	N - 1.18
228.1346	228.1386	0	0	0	A - 0.00	A - 0.00	A - 0.00	70	91	72.53	N - 3.56	V - 1.78	G - 1.68	0	315	0	A - 0.00	A - 0.00	A - 0.00
58.0226	58.0266	41	178	10.67	M - 3.41	H - 3.12	I - 2.62	35	173	4.62	M - 3.25	G - 3.05	V - 2.30	11	26	7.69	G - 7.36	H - 5.23	D - 1.54
-226.171	-226.167	77	83	72.29	I - 11.17	L - 3.21	N - 2.20	224	232	89.66	I - 10.61	L - 3.57	Q - 2.33	37	42	80.95	L - 6.52	I - 6.01	A - 0.42
71.0352	71.0392	18	33	21.21	G - 4.51	S - 2.64	K - 1.95	263	424	29.95	K - 4.13	R - 3.12	G - 2.69	78	168	14.29	R - 6.67	G - 3.03	M - 1.47
-488.229	-488.225	0	2	0	A - 0.00	A - 0.00	A - 0.00	18	25	64	Y - 26.59	P - 2.19	M - 0.93	255	256	99.61	Y - 36.74	P - 0.35	I - 0.34
-147.071	-147.067	46	51	76.47	F - 18.49	P - 1.37	S - 0.84	470	493	89.66	F - 15.09	N - 2.67	G - 1.03	78	85	80	F - 21.55	Y - 1.31	T - 0.75
185.1142	185.1182	17	30	43.33	I - 4.59	A - 3.78	Y - 2.25	774	1272	52.2	I - 5.33	D - 2.39	L - 2.18	25	29	68.97	A - 4.15	M - 3.01	Y - 2.53
12.0336	12.0376	28	38	60.53	T - 7.26	F - 5.17	G - 3.25	375	439	44.19	T - 7.71	F - 6.49	M - 1.79	120	127	38.58	F - 18.30	T - 4.00	E - 1.68
-362.165	-362.161	0	0	0	A - 0.00	A - 0.00	A - 0.00	16	22	45.45	M - 10.51	F - 7.01	S - 2.72	238	239	1.26	M - 30.92	K - 5.55	Y - 0.08
-0.049	-0.045	2	321	0.31	I - 11.70	P - 3.94	E - 3.52	55	1574	0.76	N - 2.56	A - 2.12	V - 1.66	6	465	0.22	Q - 5.22	F - 4.67	L - 2.53
-200.081	-200.077	14	14	64.29	E - 5.37	A - 3.91	Y - 2.73	85	99	39.39	M - 14.89	D - 4.20	A - 3.05	196	206	81.55	A - 5.05	E - 4.66	Y - 3.19

38.9468	38.9508	28	63	0	D - 2.75	G - 2.27	E - 1.84	1307	1409	1.99	D - 2.66	E - 1.70	V - 1.49	8	8	0	C - 6.71	D - 3.90	P - 2.85
-288.127	-288.123	4	6	0	C - 39.13	Y - 4.79	A - 0.00	93	109	5.5	C - 35.58	K - 3.31	Y - 0.21	211	220	1.36	C - 36.79	K - 2.92	E - 0.06
-225.149	-225.145	34	35	8.57	K - 9.30	P - 6.18	D - 1.23	367	398	1.51	P - 8.08	K - 6.22	D - 0.93	130	132	0.76	P - 9.03	K - 5.81	D - 0.78
46.0394	46.0434	70	99	7.07	C - 33.70	T - 0.75	D - 0.54	288	362	10.5	C - 34.30	A - 0.55	H - 0.50	3	3	66.67	C - 29.81	A - 4.77	A - 0.00
-242.14	-242.136	3	10	10	N - 17.26	K - 4.88	L - 1.12	144	220	43.18	N - 9.61	K - 4.36	G - 3.29	128	206	34.95	N - 18.44	K - 4.87	M - 0.95
-256.156	-256.152	9	14	28.57	K - 12.69	Q - 2.90	P - 0.88	346	412	54.13	K - 6.66	Q - 5.50	G - 2.40	178	201	41.79	W - 8.43	K - 5.04	Q - 4.80
213.121	213.125	15	17	64.71	E - 3.13	F - 3.12	V - 2.97	765	777	94.85	F - 5.18	V - 2.60	E - 1.98	134	139	92.09	V - 5.35	F - 5.27	E - 2.39
-127.095	-127.091	17	75	1.33	W - 4.73	K - 4.39	Y - 3.38	648	1751	5.37	P - 4.17	K - 2.25	Q - 1.72	5	33	0	I - 9.36	F - 5.61	Y - 3.83
47.9828	47.9868	18	44	18.18	W - 35.15	M - 7.53	Y - 4.25	96	172	4.07	P - 6.04	G - 4.08	W - 3.25	65	155	10.32	W - 38.87	M - 3.58	C - 1.71
-144.092	-144.088	14	30	20	M - 15.55	K - 7.53	F - 3.00	169	399	4.01	M - 22.85	F - 3.25	K - 2.31	64	107	5.61	M - 29.60	F - 1.75	K - 1.03
-259.138	-259.134	5	8	50	M - 18.34	K - 7.03	P - 3.15	114	214	33.18	M - 24.46	K - 5.68	P - 1.52	90	181	26.52	M - 24.62	K - 6.31	P - 1.63
30.0086	30.0126	37	53	3.77	A - 6.92	G - 4.11	P - 1.19	414	748	2.81	A - 6.82	G - 2.88	N - 1.20	90	112	7.14	A - 6.59	M - 2.02	F - 1.65
0.957	0.961	4	292	0.34	I - 8.77	P - 2.63	L - 2.53	115	1856	0.59	N - 3.38	I - 2.52	V - 2.43	18	551	0.54	M - 17.83	N - 3.66	I - 3.25
1.9564	1.9604	1	277	0.36	L - 10.11	A - 0.00	A - 0.00	99	1509	1.59	N - 3.58	L - 2.36	I - 1.71	5	220	0	V - 5.56	F - 5.14	D - 3.85
-369.24	-369.236	25	29	20.69	Q - 7.17	K - 6.03	L - 2.66	324	381	11.29	Q - 7.46	K - 4.04	L - 2.30	90	102	14.71	K - 4.51	Q - 2.81	L - 2.38
-228.113	-228.109	46	52	71.15	D - 6.46	I - 4.15	L - 2.20	127	139	85.61	D - 5.48	I - 2.90	V - 2.46	44	66	43.94	D - 3.86	F - 3.19	I - 1.68
17.9962	18.0002	23	169	0	M - 9.57	W - 3.49	V - 1.71	310	1035	2.13	M - 5.26	W - 4.49	N - 2.07	0	8	0	A - 0.00	A - 0.00	A - 0.00
26.0136	26.0176	7	14	28.57	A - 6.94	G - 2.17	E - 1.81	746	976	73.16	A - 2.42	V - 1.88	F - 1.63	74	88	67.05	A - 4.54	G - 2.32	F - 2.21
-14.9934	-14.9894	12	60	5	M - 9.55	H - 6.92	A - 2.98	855	2167	5.49	M - 24.88	P - 1.00	D - 0.91	0	5	0	A - 0.00	A - 0.00	A - 0.00
272.1254	272.1294	1	1	100	D - 10.50	Y - 9.57	S - 2.97	55	62	69.35	Y - 3.48	I - 2.57	V - 2.09	1	255	0.39	L - 10.11	A - 0.00	A - 0.00
253.1514	253.1554	18	18	61.11	N - 5.36	F - 4.67	E - 3.13	601	611	87.89	N - 5.58	D - 2.90	E - 1.60	135	137	78.83	E - 4.46	Q - 2.93	D - 2.50
22.9806	22.9846	46	56	0	T - 2.57	P - 1.64	N - 1.57	2924	3066	0.46	D - 2.36	E - 1.95	A - 1.43	11	16	0	E - 4.19	G - 1.95	S - 1.55
285.142	285.146	27	33	72.73	C - 4.97	Q - 4.38	I - 3.32	494	540	86.3	D - 3.37	Q - 2.33	I - 1.98	96	101	78.22	D - 6.72	I - 4.02	E - 2.45
-215.093	-215.089	19	19	94.74	S - 2.99	Q - 2.75	G - 2.49	182	217	69.12	S - 3.36	G - 2.64	H - 2.35	130	132	96.97	G - 3.58	A - 3.35	S - 2.99
248.1966	248.2006	0	1	0	A - 0.00	A - 0.00	A - 0.00	0	2	0	A - 0.00	A - 0.00	A - 0.00	0	275	0	A - 0.00	A - 0.00	A - 0.00
45.0088	45.0128	11	131	2.29	C - 4.06	P - 2.14	A - 2.02	551	2254	8.3	V - 2.31	Q - 1.44	I - 1.44	0	2	0	A - 0.00	A - 0.00	A - 0.00
188.1032	188.1072	1	1	0	R - 5.88	V - 5.56	S - 3.95	5	29	10.34	A - 4.37	I - 2.90	D - 2.60	1	234	0.43	T - 18.21	A - 0.00	A - 0.00
-186.103	-186.099	37	43	65.12	F - 4.79	V - 4.58	S - 3.27	86	135	22.22	V - 4.06	D - 3.32	S - 2.35	37	95	8.42	F - 4.19	D - 3.47	V - 3.19
-214.133	-214.129	55	56	62.5	T - 4.63	I - 3.65	L - 2.38	126	149	59.06	T - 5.01	N - 3.03	L - 2.76	66	74	54.05	T - 6.23	L - 3.08	A - 1.97
257.1352	257.1392	14	21	47.62	E - 5.87	Y - 5.47	F - 2.00	782	1269	41.77	E - 2.46	R - 2.38	V - 1.73	30	36	11.11	R - 7.15	E - 3.62	A - 1.60
1.0654	1.0694	5	253	0.4	G - 3.79	V - 3.34	D - 2.10	290	1745	0.63	E - 1.68	D - 1.56	T - 1.40	5	234	0.85	I - 9.36	F - 5.61	L - 4.05
29.9716	29.9756	7	16	6.25	W - 17.22	Y - 8.20	V - 2.98	33	71	23.94	W - 9.33	V - 4.17	D - 3.34	175	180	29.44	V - 10.81	D - 4.32	C - 2.06
159.931	159.935	13	42	0	S - 4.64	D - 3.25	P - 2.09	11	66	0	S - 4.00	T - 2.53	D - 1.93	47	120	0	S - 5.53	D - 2.91	T - 1.93
199.1296	199.1336	8	15	46.67	F - 8.76	N - 3.53	V - 3.13	793	1263	45.45	D - 2.51	V - 2.01	R - 1.89	24	27	77.78	F - 11.10	L - 3.58	Y - 2.31
-370.224	-370.22	3	5	20	V - 5.56	L - 4.50	T - 2.02	85	116	34.48	V - 5.57	K - 2.26	L - 1.77	135	167	8.98	V - 13.76	L - 1.26	K - 0.25
-343.188	-343.184	25	25	100	Q - 10.31	S - 5.85	K - 0.23	137	150	84.67	Q - 8.90	S - 5.16	K - 0.66	156	161	72.67	Q - 6.82	S - 4.93	A - 1.74
-270.171	-270.167	9	17	11.76	A - 11.12	K - 1.95	L - 1.12	28	42	7.14	K - 12.66	A - 2.89	N - 0.84	151	160	6.25	A - 12.06	K - 1.82	Y - 0.30
230.0704	230.0744	0	24	0	A - 0.00	A - 0.00	A - 0.00	2	7	28.57	Y - 19.14	H - 3.83	T - 1.82	2	187	0.53	M - 22.92	Y - 9.57	E - 3.52
81.9692	81.9732	35	121	1.65	S - 3.41	P - 2.37	D - 2.31	582	1355	0.66	S - 3.40	P - 2.65	E - 1.55	0	0	0	A - 0.00	A - 0.00	A - 0.00
-116.06	-116.056	8	8	87.5	C - 13.04	A - 3.57	T - 2.09	1079	1157	16.59	C - 20.91	H - 1.40	N - 1.13	50	55	12.73	C - 16.45	T - 3.62	Y - 1.66
271.1262	271.1302	16	20	70	C - 4.19	D - 3.94	E - 2.35	511	522	72.22	Y - 6.61	K - 3.12	V - 1.87	99	99	96.97	Y - 13.76	W - 4.19	N - 2.20
116.0606	116.0646	0	0	0	A - 0.00	A - 0.00	A - 0.00	1638	1660	96.81	G - 3.57	A - 2.18	I - 2.06	1	2	0	D - 5.60	V - 3.34	I - 3.12
0.881	0.885	2	141	0	I - 11.70	P - 7.88	A - 0.00	22	782	0.13	I - 5.32	A - 2.92	L - 2.45	4	366	0.27	Y - 19.14	G - 2.21	Q - 1.74
279.9884	279.9924	2	89	0	M - 34.38	G - 3.79	A - 0.00	0	0	0	A - 0.00	A - 0.00	A - 0.00	0	0	0	A - 0.00	A - 0.00	A - 0.00
189.0434	189.0474	1	28	3.57	E - 14.09	A - 0.00	A - 0.00	3	230	1.3	N - 28.25	A - 0.00	A - 0.00	2	91	1.1	N - 14.13	Q - 10.45	A - 0.00
129.0398	129.0438	3	10	0	E - 10.33	L - 1.35	T - 1.21	114	279	5.02	K - 6.79	N - 2.06	E - 1.90	116	150	2	R - 9.73	D - 3.65	A - 1.39
-342.228	-342.224	11	12	83.33	K - 6.79	I - 5.67	T - 5.38	85	118	54.24	T - 5.95	K - 5.57	I - 4.86	140	152	25.66	K - 6.15	I - 3.76	L - 3.59
162.0504	162.0544	0	14	0	A - 0.00	A - 0.00	A - 0.00	146	845	2.72	K - 6.53	W - 4.21	D - 1.86	36	78	3.85	W - 9.01	C - 9.00	K - 2.95
-312.218	-312.214	19	19	78.95	K - 5.40	Q - 3.30	V - 3.22	137	170	57.06	K - 5.12	A - 4.06	L - 2.49	122	141	21.99	L - 3.91	K - 3.88	A - 3.80
-1.0698	-1.0658	0	243	0	A - 0.00	A - 0.00	A - 0.00	21	1432	0.14	V - 4.06	I - 3.44	N - 3.03	3	196	0	S - 8.90	C - 3.73	Y - 3.19
-96.1078	-96.1038	45	57	0	W - 65.32	K - 1.27	C - 0.99	9	29	0	W - 37.49	I - 7.80	V - 2.22	46	62	0	W - 44.83	Y - 3.96	I - 3.05
-260.155	-260.151	40	48	79.17	F - 13.61	L - 4.57	I - 0.78	127	258	46.9	F - 12.28	L - 4.26	C - 0.82	59	73	73.97	F - 11.50	L - 4.05	C - 1.33
-0.961	-0.957	3	310	0.32	Q - 6.96	E - 6.26	F - 3.12	76	1799	1.11	N - 3.08	Q - 1.92	F - 1.89	1	107	0.93	F - 14.02	L - 5.06	A - 0.00
-75.1842	-75.1802	0	0	0	A - 0.00	A - 0.00	A - 0.00	6	44	0	M - 3.03	D - 2.28	H - 2.13	116	242	1.24	D - 3.32	N - 1.97	Y - 1.37
-289.075	-289.071	0	0	0	A - 0.00	A - 0.00	A - 0.00	85	86	93.02	C - 26.92	E - 5.14	H - 0.30	162	162	96.3	C - 23.89	E - 6.21	W - 0.25
156.0294	156.0334	0	0	0	A - 0.00	A - 0.00	A - 0.00	5	10	10	K - 3.51	Y - 3.45	G - 2.23	0	205	0	A - 0.00	A - 0.00	A - 0.00
229.1404	229.1444	6	17	29.41	V - 6.49	P - 3.50	G - 2.61	868	1203	46.38	R - 3.23	D - 2.50	E - 1.98	3	6	16.67	I - 7.80	A - 6.67	Q - 1.39
174.0864	174.0904	1	1	0	I - 11.70	S - 5.93	A - 0.00	4	21	9.52	W - 12.05	P - 1.86	Q - 1.52	0	197	0	A - 0.00	A - 0.00	A - 0.00
252.9778	252.9818	17	84	10.71	D - 5.56	T - 4.82	A - 3.78	1	2	50	V - 3.58	Q - 2.98	A - 2.04	0	2	0	A - 0.00	A - 0.00	A - 0.00
256.1884	256.1924	21	23	17.39	D - 4.83	N - 4.04	V - 3.84	1004	1113	67.21	E - 3.93	D - 2.81	V - 1.40	3	5	20	W - 13.39	E - 4.70	D - 3.50
200.1034	200.1074	2	4	25	Q - 5.22	F - 4.67	P - 3.94	33	50	44	E - 2.26	D - 1.96	C - 1.72	0	213	0	A - 0.00	A - 0.00	A - 0.00
262.139	262.143	0	1	0	A - 0.00	A - 0.00	A - 0.00	1	18	5.56	N - 5.65	F - 5.61	T - 3.64	1	213	0	Q - 20.89	A - 0.00	A - 0.00
168.0764	168.0804	1	1	100	G - 15.18	A - 0.00	A - 0.00	4	9	44.44	C - 16.77	H - 4.79	Y - 4.79	0	196	0	A - 0.00	A - 0.00	A - 0.00
70.9984	71.0024	14	26	34.62	W - 14.35	I - 6.69	K - 2.51	22	81	18.52	W - 5.76	V - 2.71	M - 2.40	3	120	0	Q - 6.96	L - 2.81	V - 2.78
152.9952	152.9992	3	3	0	C - 37.26	T - 3.03	A - 0.00	1458	1731	9.36	C - 31.71	E - 0.51	Y - 0.50	0	3	0	A - 0.00	A - 0.00	A - 0.00
30.9812	30.9852	0	6	0	A - 0.00	A - 0.00	A - 0.00	12	136	0.74	N - 3.26	G - 2.33	R - 1.90	2	219	0	W - 20.09	G - 3.79	E - 3.52
1.941	1.945	5	242	1.24	I - 9.36	K - 3.51	L - 2.02	50	1025	2.44	I - 8.94	M - 2.60	V - 1.42	19	264	5.3	I - 10.88	L - 2.75	S - 1.04
130.0984	130.1024	16	29	13.79	D - 3.49	G - 3.42	M - 2.87	1343	2049	22.94	D - 1.78	E - 1.74	Y - 1.37	0	1	0	A - 0.00	A - 0.00	A - 0.00
242.1362	242.1402	8	18	33.33	E - 3.96	D - 3.94	I - 2.92	661	910	60	E - 2.47	D - 2.06	Y - 1.95	13	16	56.25	L - 6.82	R - 1.36	G - 0.91
-400.191	-400.187	0	10	0	A - 0.00	A - 0.00	A - 0.00	20	107	11.21	R - 5.44	K - 1.87	E - 1.61	0	172	0	A - 0.00	A - 0.00	A - 0.00
42.0452	42.0492	18	46	10.87	S - 3.63	K - 2.68	A - 2.31	194	291	24.05	K - 5.04	G - 2.76	S - 2.07	42	49	22.45	G - 5.33	K - 3.47	S - 2.53
-259.102	-259.098	3	4	75	C - 14.90	V - 5.56	M - 5.09	122	130	88.46	C - 18.51	V - 7.06	M - 4.88	143	145	93.1	C - 13.49	M - 8.09	V - 4.60
216.1354	216.1394	0	0	0	A - 0.00	A - 0.00	A - 0.00	24	46	36.96	D - 6.15	V - 2.43	Q - 1.77	0	220	0	A - 0.00	A - 0.00	A - 0.00
-275.166	-275.162	6	6	33.33	F - 23.37	K - 2.93	A - 0.00	88	131	41.22	F - 16.28	K - 5.64	M - 0.73	149	162	50.62	F - 13.42	Y - 7.79	K - 4.91

-356.208	-356.204	27	29	44.83	D - 7.77	L - 4.18	K - 1.95	209	226	71.24	D - 7.30	L - 3.53	I - 1.58	85	88	73.86	D - 6.32	K - 6.27	L - 2.61
-227.141	-227.137	7	7	71.43	R - 5.55	A - 5.51	V - 3.58	786	834	79.5	R - 8.56	V - 6.73	A - 1.10	38	40	17.5	R - 10.91	A - 3.29	W - 1.06
255.1676	255.1716	10	11	63.64	I - 4.68	Q - 3.13	E - 2.82	630	641	92.2	I - 3.50	N - 2.04	H - 2.04	58	58	84.48	W - 14.31	Q - 4.32	C - 3.98
185.039	185.043	30	89	1.12	Y - 6.36	K - 4.40	V - 2.17	501	813	2.71	Y - 11.68	P - 1.30	G - 1.17	1	3	0	N - 8.07	T - 2.60	R - 2.52
99.0664	99.0704	10	34	5.88	K - 9.23	N - 3.53	D - 3.15	142	261	0.38	K - 10.32	C - 1.78	R - 1.41	9	33	18.18	M - 10.19	L - 5.43	K - 1.95
-291.16	-291.156	5	5	60	Y - 26.80	K - 5.27	A - 0.00	193	228	69.74	Y - 21.53	K - 6.56	P - 0.91	110	123	44.72	Y - 27.05	K - 5.03	D - 0.04
257.1466	257.1506	12	13	69.23	Q - 5.22	R - 4.41	D - 3.50	399	434	85.71	I - 7.52	D - 2.11	L - 1.43	100	100	97	E - 9.89	D - 3.46	G - 1.21
306.166	306.17	0	0	0	A - 0.00	A - 0.00	A - 0.00	6	9	22.22	I - 3.90	M - 3.82	N - 2.51	0	182	0	A - 0.00	A - 0.00	A - 0.00
199.1048	199.1088	58	58	86.21	G - 3.00	F - 2.82	V - 2.38	196	233	67.38	I - 3.00	M - 1.87	V - 1.79	0	3	0	A - 0.00	A - 0.00	A - 0.00
226.039	226.043	38	58	20.69	Y - 20.19	K - 3.31	E - 0.87	211	333	17.72	Y - 20.32	E - 2.83	K - 0.82	0	1	0	A - 0.00	A - 0.00	A - 0.00
-457.256	-457.252	5	7	57.14	D - 6.30	L - 5.06	T - 1.82	67	97	48.45	D - 4.96	L - 2.83	E - 2.21	114	115	89.57	D - 7.67	L - 3.90	T - 2.42
-340.25	-340.246	12	15	40	V - 10.77	L - 2.32	I - 1.95	176	234	67.95	V - 7.47	L - 4.02	I - 1.72	81	86	34.88	V - 5.53	L - 4.54	K - 3.87
115.0252	115.0292	19	28	10.71	R - 5.42	N - 3.32	G - 2.57	247	334	8.98	K - 6.39	R - 3.92	N - 2.19	43	50	20	S - 4.00	N - 3.94	K - 2.86
-355.224	-355.22	21	29	65.52	N - 5.83	Q - 3.98	K - 3.63	120	145	62.07	N - 5.52	I - 3.43	Q - 2.83	60	65	33.85	N - 6.23	I - 4.50	A - 2.18
243.131	243.135	26	26	92.31	L - 3.71	D - 3.09	F - 2.52	611	650	80.92	D - 4.33	G - 1.89	E - 1.60	52	53	96.23	D - 6.86	V - 2.83	F - 1.98
-15.0234	-15.0194	35	77	22.08	Q - 4.64	C - 4.05	H - 3.48	915	1609	35.43	Q - 4.04	C - 3.23	H - 1.90	2	15	0	E - 14.09	A - 0.00	A - 0.00
242.1246	242.1286	6	23	26.09	N - 28.25	A - 0.00	A - 0.00	58	160	28.13	N - 20.62	K - 2.73	S - 0.51	23	61	0	K - 17.57	A - 0.00	A - 0.00
39.9498	39.9538	11	18	0	D - 4.46	Q - 2.20	G - 1.76	571	687	1.89	D - 2.31	E - 1.79	A - 1.61	79	107	2.8	M - 14.62	P - 2.43	A - 2.21
147.0332	147.0372	5	45	11.11	M - 9.17	H - 7.67	R - 3.53	5	53	3.77	H - 10.22	C - 8.94	M - 4.58	3	38	7.89	A - 12.71	T - 2.02	A - 0.00
377.0432	377.0472	0	1	0	A - 0.00	A - 0.00	A - 0.00	0	1140	0	A - 0.00	A - 0.00	A - 0.00	0	0	0	A - 0.00	A - 0.00	A - 0.00
-298.202	-298.198	14	15	80	K - 5.02	V - 4.17	A - 3.57	157	372	31.72	K - 6.01	V - 3.08	G - 2.32	68	80	8.75	I - 6.59	K - 4.76	V - 3.76
312.2	312.204	31	33	63.64	D - 9.25	E - 1.64	G - 1.32	269	273	90.11	E - 3.11	M - 2.84	Q - 1.92	37	37	54.05	E - 3.71	D - 2.67	F - 2.05
163.0614	163.0654	4	12	0	K - 9.89	Y - 7.18	P - 1.97	103	228	1.32	Y - 5.75	K - 5.67	R - 3.57	18	100	1	R - 9.29	M - 7.02	Y - 3.26
311.1278	311.1318	42	59	64.41	A - 2.72	F - 2.67	G - 2.29	43	72	54.17	Y - 4.90	V - 4.65	D - 2.66	1	1	0	D - 10.50	T - 9.10	A - 0.00
300.1188	300.1228	0	1	0	A - 0.00	A - 0.00	A - 0.00	4	142	1.41	V - 5.14	C - 4.47	M - 2.29	1	170	0	R - 17.65	A - 0.00	A - 0.00
-2.0546	-2.0506	8	160	3.13	N - 6.47	I - 5.36	W - 5.02	67	973	4.11	L - 3.46	N - 2.32	I - 1.83	2	56	1.79	I - 5.85	A - 3.57	V - 2.09
18.03	18.034	7	125	0	W - 11.48	A - 4.09	T - 3.53	89	1225	0.24	I - 1.99	Y - 1.64	L - 1.62	0	12	0	A - 0.00	A - 0.00	A - 0.00
-1.079	-1.075	2	224	0	F - 7.01	D - 3.94	S - 2.97	33	1358	0.74	A - 2.99	I - 2.78	V - 2.28	1	201	0	A - 14.30	A - 0.00	A - 0.00
282.1666	282.1706	0	20	0	A - 0.00	A - 0.00	A - 0.00	8	154	0	K - 14.28	F - 3.51	I - 1.46	5	85	3.53	I - 6.76	R - 4.71	V - 2.22
-0.9962	-0.9922	6	292	0.68	H - 6.39	F - 3.12	T - 3.03	136	2374	1.68	M - 3.35	N - 2.59	Q - 1.44	8	319	0	M - 4.01	N - 3.53	V - 3.20
-242.129	-242.125	16	16	62.5	E - 5.14	I - 4.39	L - 2.05	99	113	69.91	E - 5.86	I - 4.70	L - 1.78	85	90	41.11	L - 5.37	E - 4.74	Y - 2.25
303.9904	303.9944	6	103	0	H - 5.34	T - 3.45	D - 2.34	139	176	10.8	C - 35.34	I - 0.95	D - 0.60	0	0	0	A - 0.00	A - 0.00	A - 0.00
-218.074	-218.07	18	18	83.33	M - 21.22	S - 4.83	A - 1.85	152	153	88.89	M - 15.86	S - 5.80	A - 2.22	89	89	55.06	M - 24.87	S - 5.07	A - 0.39
256.1522	256.1562	1	6	16.67	L - 10.11	A - 0.00	A - 0.00	474	878	33.94	H - 3.17	F - 2.35	R - 1.98	4	8	37.5	I - 5.85	V - 4.17	A - 3.57
-440.241	-440.237	7	7	71.43	N - 10.09	L - 5.06	K - 1.26	251	267	83.15	N - 11.99	L - 4.24	K - 0.67	86	89	94.38	N - 13.80	L - 4.94	P - 0.37
-493.267	-493.263	0	12	0	A - 0.00	A - 0.00	A - 0.00	24	122	4.92	H - 18.24	M - 1.91	I - 1.55	11	93	0	K - 12.78	H - 10.45	A - 0.00
-427.245	-427.241	8	10	40	I - 12.43	E - 3.96	V - 2.09	218	248	56.05	I - 12.06	E - 3.48	K - 1.29	77	81	43.21	I - 10.33	E - 2.26	G - 2.07
-227.129	-227.125	31	33	69.7	N - 4.40	V - 3.00	I - 2.77	64	88	60.23	N - 5.19	V - 3.43	F - 2.45	44	48	25	I - 9.84	A - 3.79	N - 1.93
259.133	259.137	4	5	60	D - 10.50	I - 2.92	Q - 2.61	536	856	53.04	E - 5.45	D - 3.25	Q - 2.07	12	22	40.91	D - 6.12	G - 3.37	Q - 3.19
-253.156	-253.152	11	11	36.36	R - 11.76	P - 4.77	V - 0.51	310	329	33.43	R - 8.98	P - 3.80	K - 2.21	88	89	49.44	R - 8.59	P - 3.13	K - 2.25
-15.0132	-15.0092	22	39	2.56	N - 15.39	M - 3.72	G - 2.41	483	873	20.85	M - 7.10	N - 5.43	Q - 2.44	46	59	8.47	M - 12.07	N - 10.34	H - 2.10
-137.061	-137.057	23	26	84.62	H - 29.16	Y - 1.11	G - 0.99	287	356	76.12	H - 24.97	E - 2.11	M - 1.60	21	21	100	H - 20.08	Y - 5.47	L - 2.41
-28.0334	-28.0294	20	25	24	V - 9.28	A - 2.71	H - 1.92	281	410	24.15	V - 9.99	D - 2.45	A - 2.01	23	47	0	V - 13.31	D - 2.45	S - 0.41
-156.092	-156.088	42	42	80.95	V - 5.61	G - 4.68	P - 1.70	84	119	56.3	V - 5.92	G - 5.07	M - 3.71	31	32	15.63	G - 9.91	V - 2.28	E - 0.86
54.938	54.942	35	51	0	Y - 19.42	F - 3.20	N - 0.81	96	189	1.06	Y - 15.10	A - 2.45	C - 1.56	3	3	0	Y - 19.14	S - 3.95	I - 3.90
462.096	462.1	0	0	0	A - 0.00	A - 0.00	A - 0.00	0	1129	0	A - 0.00	A - 0.00	A - 0.00	0	0	0	A - 0.00	A - 0.00	A - 0.00
-1.0914	-1.0874	3	195	0	L - 6.74	P - 2.63	N - 2.35	17	1121	0.36	V - 4.51	I - 2.43	L - 2.33	0	184	0	A - 0.00	A - 0.00	A - 0.00
-234.103	-234.099	13	13	92.31	Y - 8.35	F - 8.27	S - 3.50	69	79	84.81	Y - 12.29	A - 4.59	F - 3.69	90	91	61.54	Y - 8.58	L - 5.77	A - 2.94
1.9196	1.9236	2	202	0	T - 4.55	V - 4.17	P - 2.63	29	1034	0.29	T - 3.35	Q - 2.73	V - 2.49	10	209	0	Q - 4.87	L - 3.54	V - 2.22
276.1194	276.1234	1	57	0	D - 10.50	L - 5.06	A - 0.00	2	8	25	W - 28.12	D - 5.25	Q - 4.18	0	3	0	A - 0.00	A - 0.00	A - 0.00
256.1294	256.1334	0	1	0	A - 0.00	A - 0.00	A - 0.00	14	26	46.15	V - 2.58	D - 2.25	A - 2.18	1	157	0.64	E - 9.39	L - 3.37	A - 0.00
-2.0064	-2.0024	2	113	0	F - 7.01	Q - 5.22	C - 4.47	107	1102	4.63	W - 8.63	M - 8.07	V - 1.80	2	37	2.7	D - 12.24	M - 7.64	H - 3.19
-243.136	-243.132	4	7	0	S - 8.40	R - 3.68	L - 0.84	141	239	6.69	R - 8.23	S - 4.90	C - 0.74	74	81	1.23	R - 8.90	S - 5.50	A - 0.23
-258.087	-258.083	0	0	0	A - 0.00	A - 0.00	A - 0.00	6	10	30	E - 9.59	M - 1.91	D - 1.37	125	127	90.55	E - 13.13	A - 0.46	S - 0.24
26.0498	26.0538	10	13	23.08	S - 8.11	T - 2.73	M - 2.29	111	143	28.67	S - 8.16	T - 2.65	C - 1.41	49	78	10.26	S - 11.02	T - 1.30	A - 0.00
-1.022	-1.018	12	428	0.47	W - 12.27	C - 4.47	M - 3.18	189	2459	4.35	W - 6.76	M - 4.33	I - 2.04	121	268	33.58	L - 2.36	I - 2.29	A - 1.73
370.218	370.222	0	14	0	A - 0.00	A - 0.00	A - 0.00	207	366	24.04	C - 5.81	R - 3.39	E - 2.76	24	84	0	C - 21.43	R - 8.82	P - 0.33
97.9644	97.9684	2	33	0	F - 14.02	M - 11.46	S - 2.97	4	43	0	Y - 9.57	G - 2.60	P - 2.36	0	107	0	A - 0.00	A - 0.00	A - 0.00
176.7442	176.7482	0	0	0	A - 0.00	A - 0.00	A - 0.00	1	14	0	G - 10.12	Q - 6.96	A - 0.00	27	186	0	D - 7.99	E - 2.96	V - 2.31
-414.141	-414.137	1	2	0	D - 20.99	A - 0.00	A - 0.00	43	43	16.28	D - 18.06	P - 0.73	R - 0.57	118	119	0.84	D - 17.89	P - 1.27	R - 1.16
-25.034	-25.03	0	0	0	A - 0.00	A - 0.00	A - 0.00	558	1059	11.9	C - 29.96	E - 1.01	I - 0.64	7	10	20	C - 34.07	E - 1.68	P - 1.13
-214.098	-214.094	41	42	78.57	D - 7.95	V - 6.28	N - 1.15	45	52	73.08	D - 7.01	V - 6.24	N - 0.98	12	16	68.75	D - 5.98	V - 4.75	A - 1.19
176.1388	176.1428	0	0	0	A - 0.00	A - 0.00	A - 0.00	1	6	0	A - 6.50	P - 2.86	E - 2.56	0	130	0	A - 0.00	A - 0.00	A - 0.00
72.0312	72.0352	3	28	0	W - 40.17	A - 4.77	N - 4.71	105	617	2.59	W - 43.31	H - 1.03	G - 0.92	0	1	0	A - 0.00	A - 0.00	A - 0.00
123.9104	123.9144	0	0	0	A - 0.00	A - 0.00	A - 0.00	0	1	0	A - 0.00	A - 0.00	A - 0.00	22	114	0	W - 25.56	A - 3.90	T - 3.72
-241.145	-241.141	15	18	66.67	A - 5.88	V - 2.60	Q - 2.55	75	100	50	Q - 5.43	I - 2.30	L - 2.11	73	82	43.9	A - 3.90	F - 3.39	Q - 1.53
134.045	134.049	0	0	0	A - 0.00	A - 0.00	A - 0.00	1	10	10	I - 5.01	S - 2.54	Q - 1.49	0	119	0	A - 0.00	A - 0.00	A - 0.00
15.0168	15.0208	19	37	8.11	W - 4.23	H - 3.36	V - 3.07	416	995	6.33	V - 3.37	N - 2.86	H - 1.69	0	4	0	A - 0.00	A - 0.00	A - 0.00
24.9394	24.9434	0	0	0	A - 0.00	A - 0.00	A - 0.00	5	12	8.33	A - 3.81	N - 2.56	V - 1.95	118	137	2.92	N - 10.43	D - 4.55	V - 2.22
243.1206	243.1246	15	16	75	Y - 10.21	D - 5.60	Q - 2.79	492	646	67.03	Y - 3.44	F - 3.38	D - 2.57	47	56	82.14	L - 8.61	D - 1.56	G - 0.65
406.1576	406.1616	0	32	0	A - 0.00	A - 0.00	A - 0.00	4	264	0.76	C - 11.18	G - 4.74	S - 2.22	0	39	0	A - 0.00	A - 0.00	A - 0.00
66.9252	66.9292	0	0	0	A - 0.00	A - 0.00	A - 0.00	36	118	0	N - 6.99	E - 4.40	F - 3.95	126	165	0	H - 20.85	N - 2.21	G - 1.13
204.081	204.085	3	63	0	I - 3.90	A - 3.57	S - 2.97	31	406	1.23	V - 2.77	T - 2.40	N - 2.23	0	3	0	A - 0.00	A - 0.00	A - 0.00
-202.062	-202.058	17	17	100	D - 8.03	S - 5.70	V - 1.31	113	116	95.69	D - 7.49	S - 4.29	V - 1.97	61	62	98.39	D - 7.03	V - 4.38	S - 3.97

-399.214	-399.21	8	9	77.78	S - 6.92	P - 4.59	K - 2.20	60	78	64.1	S - 6.08	P - 3.19	D - 1.34	93	93	68.82	S - 5.08	P - 3.22	A - 2.46
-10.034	-10.03	0	5	0	A - 0.00	A - 0.00	A - 0.00	19	73	0	D - 5.12	H - 4.18	V - 2.44	37	103	3.88	H - 19.85	N - 2.80	T - 1.39
1.0404	1.0444	6	265	0.38	E - 4.70	L - 3.09	Q - 1.74	1683	4039	0.22	D - 1.57	E - 1.33	Y - 1.31	4	296	0	W - 10.04	Y - 4.79	Q - 3.66
345.0168	345.0208	0	0	0	A - 0.00	A - 0.00	A - 0.00	19	940	0.32	C - 41.58	Y - 1.01	D - 0.18	0	0	0	A - 0.00	A - 0.00	A - 0.00
284.9576	284.9616	6	61	8.2	K - 7.32	Q - 6.62	C - 1.86	0	3	0	A - 0.00	A - 0.00	A - 0.00	0	1	0	A - 0.00	A - 0.00	A - 0.00
-172.087	-172.083	30	32	84.38	A - 5.40	T - 5.21	I - 4.81	43	58	51.72	T - 6.01	A - 3.80	L - 3.31	28	32	28.13	F - 8.35	A - 4.55	T - 2.55
-471.252	-471.248	5	8	25	W - 20.09	V - 10.01	I - 1.17	27	34	79.41	F - 7.47	L - 3.64	W - 2.98	150	151	96.03	F - 12.39	L - 4.72	N - 0.52
-469.292	-469.288	12	12	91.67	V - 6.26	L - 2.95	K - 2.68	102	122	78.69	V - 7.07	L - 4.04	I - 1.61	74	78	79.49	I - 5.74	V - 5.60	E - 2.60
114.0416	114.0456	26	33	30.3	K - 6.53	G - 6.23	C - 6.02	102	176	10.8	K - 10.60	C - 1.75	G - 1.62	13	26	30.77	C - 27.52	K - 3.38	G - 1.75
275.1612	275.1652	7	15	20	E - 10.74	D - 5.00	A - 0.00	344	584	44.52	D - 5.34	E - 2.14	V - 1.87	8	11	9.09	D - 7.43	E - 4.40	M - 2.87
384.0796	384.0836	0	3	0	A - 0.00	A - 0.00	A - 0.00	0	924	0	A - 0.00	A - 0.00	A - 0.00	2	2	50	W - 5.74	L - 3.47	S - 2.37
186.087	186.091	0	1	0	A - 0.00	A - 0.00	A - 0.00	4	15	13.33	R - 2.21	S - 2.08	D - 1.83	2	125	0	L - 10.11	A - 0.00	A - 0.00
17.9532	17.9572	3	8	12.5	L - 7.87	I - 2.60	S - 1.32	64	272	2.21	L - 6.51	I - 5.42	D - 0.66	16	69	1.45	L - 6.64	I - 4.39	T - 1.14
-198.139	-198.135	42	42	83.33	V - 12.91	D - 1.79	L - 0.70	28	41	56.1	V - 12.31	W - 1.91	N - 1.35	2	3	66.67	V - 13.90	P - 2.63	A - 0.00
236.0658	236.0698	20	26	65.38	S - 4.97	M - 2.29	D - 1.77	111	127	49.61	S - 4.55	P - 1.84	A - 1.59	67	70	82.86	S - 5.58	A - 2.80	V - 1.33
-0.9434	-0.9394	1	233	0	W - 40.17	L - 5.06	A - 0.00	61	1251	0.64	V - 3.58	N - 2.66	E - 2.01	4	92	2.17	C - 11.18	G - 3.79	N - 3.53
105.8226	105.8266	0	0	0	A - 0.00	A - 0.00	A - 0.00	0	5	0	A - 0.00	A - 0.00	A - 0.00	21	140	0	M - 4.26	N - 4.06	Y - 3.45
284.1818	284.1858	6	36	5.56	V - 6.49	G - 3.16	K - 1.46	185	358	48.88	E - 7.50	I - 6.25	D - 0.88	8	9	88.89	D - 17.05	E - 2.64	A - 0.00
-325.165	-325.161	1	1	100	I - 7.80	D - 7.00	P - 5.25	13	18	38.89	V - 4.76	D - 3.84	I - 3.76	113	114	97.37	I - 7.18	D - 6.90	P - 5.22
359.0062	359.0102	1	11	0	M - 9.17	T - 7.28	V - 3.34	0	1	0	A - 0.00	A - 0.00	A - 0.00	1	98	0	C - 22.36	N - 14.13	A - 0.00
1.8758	1.8798	0	111	0	A - 0.00	A - 0.00	A - 0.00	8	489	0	A - 4.07	I - 3.46	V - 1.38	2	160	0	K - 8.79	I - 4.68	T - 1.82
-313.166	-313.162	4	4	100	A - 3.87	E - 3.23	I - 2.92	58	67	56.72	V - 2.97	A - 2.66	E - 2.13	117	119	33.61	A - 7.74	E - 2.38	L - 2.29
-1.0436	-1.0396	1	336	0.3	C - 11.18	L - 5.06	V - 4.17	61	1686	1.6	M - 2.76	I - 2.41	L - 2.35	5	176	1.7	I - 5.62	C - 4.47	V - 4.17
-185.118	-185.114	3	8	12.5	P - 10.50	N - 3.14	K - 1.95	135	196	26.53	N - 5.91	G - 4.55	K - 3.64	73	77	42.86	G - 7.02	K - 4.03	W - 2.48
-411.287	-411.283	10	12	75	L - 5.56	G - 4.81	K - 1.76	64	80	60	L - 4.35	G - 4.09	V - 1.86	62	63	92.06	G - 6.37	L - 5.60	I - 0.50
14.9714	14.9754	45	52	11.54	W - 69.34	G - 0.67	Y - 0.50	58	124	20.16	W - 36.71	L - 2.73	P - 1.22	2	19	0	W - 80.35	A - 0.00	A - 0.00
0.8454	0.8494	0	74	0	A - 0.00	A - 0.00	A - 0.00	3	417	0.24	L - 4.50	P - 3.50	A - 3.18	3	192	1.04	I - 10.40	E - 4.70	A - 3.18
326.1916	326.1956	0	15	0	A - 0.00	A - 0.00	A - 0.00	30	176	14.2	G - 8.85	C - 4.72	M - 2.32	1	87	0	N - 14.13	R - 8.82	A - 0.00
-30.0124	-30.0084	21	21	9.52	T - 11.25	S - 3.26	C - 1.60	232	262	12.21	T - 12.35	S - 1.38	C - 0.66	17	45	6.67	T - 10.71	S - 2.58	V - 1.47
-356.219	-356.215	8	15	13.33	R - 12.13	Q - 2.61	L - 1.90	80	147	6.8	R - 12.74	W - 1.34	L - 1.00	50	56	0	R - 11.14	L - 2.54	V - 1.39
-410.291	-410.287	4	5	80	I - 8.77	L - 4.64	V - 2.78	89	94	93.62	I - 9.99	L - 4.58	A - 1.20	70	72	97.22	I - 10.86	L - 4.91	V - 0.48
-13.992	-13.988	4	14	7.14	G - 3.95	L - 2.53	M - 1.91	585	1649	0.79	M - 23.29	P - 1.02	D - 0.91	1	2	0	M - 45.84	A - 0.00	A - 0.00
272.1612	272.1652	0	0	0	A - 0.00	A - 0.00	A - 0.00	36	47	68.09	Y - 6.55	W - 3.51	G - 2.57	0	128	0	A - 0.00	A - 0.00	A - 0.00
-256.119	-256.115	11	13	76.92	Q - 18.04	G - 0.69	A - 0.65	49	70	61.43	Q - 14.51	S - 1.39	H - 0.75	59	59	100	N - 9.22	A - 8.85	Q - 1.15
-358.188	-358.184	10	12	41.67	L - 6.32	E - 3.52	T - 1.37	123	193	56.48	T - 7.27	E - 6.29	L - 0.63	39	49	0	E - 12.29	L - 0.83	K - 0.41
0.0708	0.0748	2	203	0	V - 8.34	T - 4.55	A - 3.57	186	1466	0.89	M - 1.79	I - 1.73	L - 1.45	3	173	0.58	F - 7.01	E - 4.70	G - 3.79
355.902	355.906	18	61	1.64	P - 2.49	G - 2.11	D - 1.75	0	0	0	A - 0.00	A - 0.00	A - 0.00	0	0	0	A - 0.00	A - 0.00	A - 0.00
-160.033	-160.029	12	12	91.67	C - 27.95	Y - 3.19	V - 1.39	229	236	77.97	C - 28.45	V - 1.89	A - 0.70	45	49	89.8	C - 24.26	L - 1.85	S - 1.27
-184.087	-184.083	5	5	80	D - 5.25	P - 3.94	S - 3.56	39	48	52.08	W - 8.24	P - 4.21	S - 3.17	89	100	31	D - 12.13	P - 2.08	N - 1.53
-1.9696	-1.9656	0	135	0	A - 0.00	A - 0.00	A - 0.00	43	1121	1.16	Y - 3.34	I - 3.18	V - 3.06	5	52	0	D - 16.79	L - 2.02	A - 0.00
-399.251	-399.247	12	12	91.67	S - 5.44	L - 4.64	D - 0.44	76	106	58.49	S - 4.34	L - 3.80	A - 1.17	57	61	9.84	S - 7.09	I - 2.46	K - 1.49
-17.0108	-17.0068	49	70	18.57	D - 2.77	T - 2.09	E - 1.89	608	799	33.17	D - 2.79	Q - 2.44	N - 1.74	25	44	22.73	Q - 3.61	D - 1.91	A - 1.85
-441.296	-441.292	2	3	33.33	T - 9.10	I - 5.85	V - 4.17	28	63	22.22	L - 3.71	I - 3.46	A - 2.81	82	86	37.21	I - 7.66	L - 3.10	V - 3.09
-244.071	-244.067	12	12	66.67	D - 10.50	E - 5.87	A - 1.19	72	83	81.93	D - 10.25	E - 5.79	S - 0.66	66	73	89.04	D - 7.81	E - 7.32	S - 0.59
-213.15	-213.146	0	1	0	A - 0.00	A - 0.00	A - 0.00	24	30	50	Q - 7.40	K - 5.06	A - 2.78	103	105	7.62	I - 7.49	K - 5.48	P - 5.14
36.917	36.921	0	0	0	A - 0.00	A - 0.00	A - 0.00	4	29	0	M - 13.37	C - 5.59	Q - 3.48	85	149	0	M - 12.26	D - 2.30	Q - 1.61
-318.139	-318.135	0	9	0	A - 0.00	A - 0.00	A - 0.00	0	63	0	A - 0.00	A - 0.00	A - 0.00	0	52	0	A - 0.00	A - 0.00	A - 0.00
-328.177	-328.173	1	7	14.29	Q - 10.45	L - 5.06	A - 0.00	26	42	23.81	Q - 6.16	E - 2.30	I - 1.95	76	85	32.94	E - 3.37	N - 3.35	K - 1.85
275.128	275.132	26	34	76.47	D - 12.92	Q - 2.28	L - 2.07	78	101	59.41	E - 4.91	D - 4.31	T - 3.31	1	2	50	G - 15.18	A - 0.00	A - 0.00
414.2442	414.2482	0	7	0	A - 0.00	A - 0.00	A - 0.00	35	139	16.55	Y - 6.20	L - 2.44	D - 1.93	0	77	0	A - 0.00	A - 0.00	A - 0.00
225.1462	225.1502	6	7	57.14	V - 8.34	E - 3.52	D - 3.50	90	106	54.72	W - 3.57	D - 2.99	F - 2.83	72	72	100	W - 65.84	I - 2.49	V - 0.93
-413.23	-413.226	18	19	52.63	D - 7.00	T - 3.20	P - 2.63	67	102	39.22	D - 3.68	T - 2.94	V - 1.74	44	49	71.43	T - 4.19	Q - 2.30	D - 2.27
-414.214	-414.21	8	10	20	L - 7.59	V - 1.74	A - 1.04	47	65	24.62	L - 5.30	D - 2.57	V - 1.99	26	64	21.88	I - 3.90	A - 3.41	L - 3.21
-264.114	-264.11	11	11	100	Y - 16.53	T - 7.86	D - 1.43	141	145	97.24	Y - 18.35	T - 8.42	A - 0.60	57	57	73.68	Y - 22.78	T - 6.36	A - 0.73
-273.116	-273.112	8	8	100	C - 10.87	W - 10.04	I - 4.71	119	123	95.93	W - 15.87	C - 10.77	S - 2.44	84	87	89.66	C - 18.36	L - 3.79	Y - 3.65
239.1248	239.1288	17	18	94.44	F - 6.60	V - 2.94	L - 2.38	550	581	93.98	V - 2.46	Y - 2.21	H - 1.92	2	2	100	V - 16.68	A - 0.00	A - 0.00
291.1562	291.1602	2	6	33.33	D - 20.99	A - 0.00	A - 0.00	273	444	47.07	E - 4.61	D - 2.29	Q - 1.44	28	28	100	D - 20.99	A - 0.00	A - 0.00
-414.261	-414.257	12	12	0	T - 13.65	G - 3.79	A - 0.00	35	39	0	T - 15.08	G - 2.17	S - 0.34	62	63	0	T - 14.34	G - 3.14	K - 0.09
58.0034	58.0074	12	21	0	A - 13.11	M - 0.76	G - 0.51	189	263	6.46	G - 10.13	K - 2.42	A - 1.14	13	40	7.5	G - 5.06	N - 4.13	M - 3.53
-32.0096	-32.0056	8	14	14.29	M - 36.29	C - 5.59	A - 0.60	9	36	0	M - 18.21	Q - 1.97	F - 1.56	37	76	6.58	M - 36.39	E - 0.93	L - 0.68
-173.108	-173.104	10	11	90.91	S - 4.94	L - 4.21	T - 3.03	7	11	45.45	W - 3.35	P - 2.57	T - 1.73	55	56	96.43	T - 5.74	S - 4.06	L - 3.46
-89.0316	-89.0276	10	18	55.56	M - 34.38	T - 1.52	D - 1.40	185	251	71.71	M - 19.13	A - 4.80	D - 0.92	26	34	67.65	M - 31.72	P - 1.21	A - 0.91
-97.0554	-97.0514	15	15	26.67	P - 11.03	K - 2.93	F - 1.87	107	120	18.33	P - 11.21	M - 4.00	K - 2.93	41	41	7.32	P - 8.20	K - 6.64	W - 5.88
366.0678	366.0718	2	5	0	Y - 38.29	A - 0.00	A - 0.00	31	238	3.36	Y - 22.50	F - 1.99	I - 1.51	0	97	0	A - 0.00	A - 0.00	A - 0.00
-1.9338	-1.9298	1	127	0	V - 16.68	A - 0.00	A - 0.00	33	764	2.23	M - 3.66	I - 3.64	E - 2.50	0	24	0	A - 0.00	A - 0.00	A - 0.00
-314.161	-314.157	9	11	63.64	G - 5.06	K - 3.91	E - 3.13	101	133	63.16	G - 3.60	S - 2.12	K - 1.74	48	65	44.62	K - 5.37	G - 4.53	N - 3.14
-386.182	-386.178	0	2	0	A - 0.00	A - 0.00	A - 0.00	49	98	25.51	A - 4.57	D - 3.43	E - 2.01	71	76	26.32	A - 8.88	P - 1.11	S - 1.01
496.3094	496.3134	0	0	0	A - 0.00	A - 0.00	A - 0.00	45	50	78	V - 6.88	C - 4.14	M - 2.21	88	88	100	V - 16.59	L - 0.06	A - 0.00
284.1586	284.1626	3	3	100	E - 10.57	S - 1.98	I - 1.95	253	267	84.27	E - 5.77	Y - 2.42	S - 1.40	68	72	81.94	G - 6.46	W - 4.14	E - 3.81
-29.9942	-29.9902	10	12	0	M - 43.55	P - 0.79	A - 0.00	87	110	20.91	M - 29.97	Y - 3.91	D - 0.84	67	68	2.94	M - 19.67	Y - 7.86	P - 3.94
-112.081	-112.077	12	19	31.58	I - 4.39	L - 3.59	A - 1.20	186	344	26.45	W - 9.17	L - 4.04	I - 2.53	34	39	74.36	W - 24.69	H - 6.14	Q - 3.35
-12.0376	-12.0336	1	3	0	N - 14.13	T - 9.10	A - 0.00	118	170	6.47	I - 8.92	G - 2.24	F - 2.04	122	126	0	I - 18.25	D - 1.19	G - 1.00
-435.164	-435.16	0	0	0	A - 0.00	A - 0.00	A - 0.00	2	12	8.33	M - 11.46	K - 4.39	Y - 2.73	80	81	0	M - 28.37	K - 6.70	A - 0.00
-1.1238	-1.1198	0	103	0	A - 0.00	A - 0.00	A - 0.00	3	623	0	V - 11.12	N - 4.71	A - 2.38	1	173	0	V - 16.68	A - 0.00	A - 0.00

-185.082	-185.078	24	30	63.33	G - 3.73	Y - 2.92	P - 2.63	17	60	13.33	A - 5.68	N - 3.60	G - 2.23	13	13	92.31	Q - 6.16	G - 4.48	Y - 2.95
212.1394	212.1434	0	0	0	A - 0.00	A - 0.00	A - 0.00	5	23	17.39	F - 6.17	M - 2.84	T - 2.05	3	113	0	F - 8.57	D - 6.41	T - 5.56
-213.113	-213.109	20	21	57.14	N - 7.53	V - 6.33	T - 2.20	28	40	42.5	N - 7.23	V - 4.61	A - 3.99	32	32	31.25	A - 11.51	N - 2.50	V - 1.48
0.0822	0.0862	4	185	1.08	V - 4.17	E - 4.11	S - 2.47	93	1121	0.98	W - 4.00	F - 2.11	M - 1.91	5	131	1.53	M - 7.64	Q - 4.18	P - 3.15
-186.081	-186.077	31	31	96.77	W - 47.52	F - 2.71	N - 1.37	105	108	89.81	W - 50.09	N - 2.56	F - 1.82	18	18	94.44	W - 43.04	A - 3.97	E - 0.78
230.1132	230.1172	0	0	0	A - 0.00	A - 0.00	A - 0.00	1	21	0	N - 14.13	R - 8.82	A - 0.00	2	138	0	A - 9.53	P - 2.63	E - 2.35
214.1186	214.1226	1	3	33.33	I - 11.70	S - 5.93	A - 0.00	41	47	72.34	F - 3.25	D - 1.99	V - 1.83	0	142	0	A - 0.00	A - 0.00	A - 0.00
184.1306	184.1346	29	31	77.42	G - 13.27	P - 0.90	Y - 0.66	149	157	66.24	G - 10.89	Y - 2.57	A - 1.39	0	1	0	A - 0.00	A - 0.00	A - 0.00
-471.282	-471.278	15	17	0	Q - 11.14	K - 1.95	E - 1.57	69	80	11.25	Q - 13.37	S - 1.36	K - 1.06	49	50	0	Q - 14.92	K - 1.63	E - 1.25
208.0596	208.0636	9	23	8.7	N - 3.66	T - 2.85	S - 2.31	83	117	39.32	S - 3.60	E - 2.17	P - 2.02	17	18	44.44	D - 3.23	F - 2.77	E - 2.15
190.155	190.159	0	0	0	A - 0.00	A - 0.00	A - 0.00	2	115	1.74	I - 11.70	Y - 6.38	T - 3.03	0	85	0	A - 0.00	A - 0.00	A - 0.00
2.9508	2.9548	6	82	1.22	W - 13.39	V - 3.71	Q - 3.48	48	377	0.27	N - 4.16	Q - 2.48	M - 2.01	2	36	0	N - 28.25	A - 0.00	A - 0.00
99.0298	99.0338	0	12	0	A - 0.00	A - 0.00	A - 0.00	5	98	5.1	Q - 8.36	G - 4.55	M - 3.06	4	50	8	D - 5.25	A - 4.47	K - 4.39
303.168	303.172	3	7	42.86	N - 6.28	D - 4.66	L - 3.93	255	264	88.64	H - 4.36	I - 2.68	M - 2.18	77	81	85.19	M - 4.17	I - 3.85	D - 2.77
285.2038	285.2078	1	24	0	Q - 6.96	D - 3.50	V - 2.78	40	425	6.35	D - 3.11	E - 2.65	F - 2.48	0	11	0	A - 0.00	A - 0.00	A - 0.00
158.106	158.11	13	19	47.37	S - 2.60	H - 2.05	D - 1.99	1103	1323	53.29	E - 2.00	H - 1.64	F - 1.61	0	0	0	A - 0.00	A - 0.00	A - 0.00
160.9344	160.9384	11	30	0	S - 4.64	E - 2.51	P - 2.11	15	55	0	T - 4.24	P - 2.86	S - 1.90	0	11	0	A - 0.00	A - 0.00	A - 0.00
-354.265	-354.261	11	14	50	I - 23.40	A - 0.00	A - 0.00	80	126	38.1	I - 8.60	K - 4.63	L - 3.50	43	50	12	L - 6.04	I - 5.26	K - 3.13
-344.208	-344.204	1	1	100	D - 10.50	L - 5.06	A - 0.00	6	11	9.09	F - 10.90	S - 2.64	T - 2.63	72	76	81.58	T - 5.54	S - 3.45	L - 2.90
-16.0436	-16.0396	22	29	37.93	Q - 6.50	N - 5.19	C - 5.15	322	415	39.52	N - 8.34	C - 3.05	Q - 2.67	24	26	15.38	T - 4.56	N - 3.19	H - 2.80
-435.198	-435.194	0	0	0	A - 0.00	A - 0.00	A - 0.00	53	59	81.36	C - 22.22	F - 12.17	I - 1.03	99	102	95.1	C - 21.76	F - 13.93	Y - 0.39
458.271	458.275	0	4	0	A - 0.00	A - 0.00	A - 0.00	10	94	9.57	D - 3.87	W - 2.68	S - 2.41	0	69	0	A - 0.00	A - 0.00	A - 0.00
-282.171	-282.167	3	4	75	K - 5.86	P - 5.25	G - 5.06	47	234	14.1	G - 6.24	K - 5.92	P - 3.97	18	61	19.67	G - 8.15	P - 3.21	K - 3.09
19.0296	19.0336	6	41	0	D - 4.37	E - 2.82	T - 1.97	107	948	0	A - 2.50	C - 1.74	E - 1.63	0	5	0	A - 0.00	A - 0.00	A - 0.00
-285.171	-285.167	17	18	77.78	T - 6.87	A - 5.40	L - 1.69	32	38	76.32	T - 5.90	A - 5.06	L - 2.58	33	40	40	A - 6.77	T - 4.66	L - 1.55
-369.202	-369.198	9	11	54.55	G - 6.32	P - 3.06	A - 1.59	114	125	42.4	G - 8.70	Q - 1.30	K - 1.25	37	40	47.5	G - 8.24	L - 2.96	Q - 2.73
61.9096	61.9136	1	1	0	D - 5.25	P - 3.94	G - 3.79	2	13	0	V - 8.34	N - 4.71	D - 3.50	100	130	0	N - 6.46	E - 2.39	D - 1.77
355.904	355.908	13	53	0	G - 2.90	P - 2.71	D - 1.61	0	0	0	A - 0.00	A - 0.00	A - 0.00	0	0	0	A - 0.00	A - 0.00	A - 0.00
-99.081	-99.077	3	5	0	R - 7.84	V - 7.41	A - 1.59	62	75	6.67	R - 13.77	D - 0.80	A - 0.66	66	67	0	R - 17.38	E - 0.21	A - 0.00
-199.098	-199.094	11	15	53.33	Q - 7.28	A - 6.28	D - 1.91	59	69	53.62	A - 6.40	Q - 6.23	P - 0.80	41	44	79.55	A - 5.58	Q - 4.50	W - 1.96
-402.159	-402.155	0	0	0	A - 0.00	A - 0.00	A - 0.00	4	7	57.14	C - 9.22	E - 4.32	I - 2.92	97	98	98.98	C - 21.05	E - 4.75	L - 1.65
-300.146	-300.142	0	8	0	A - 0.00	A - 0.00	A - 0.00	30	106	21.7	D - 5.46	K - 3.87	G - 3.17	35	47	63.83	D - 6.10	K - 5.10	G - 3.76
58.9982	59.0022	26	31	77.42	M - 31.74	W - 12.36	A - 1.19	63	128	21.09	M - 15.53	C - 3.02	G - 1.88	2	3	33.33	N - 7.06	T - 4.55	H - 3.83
-271.155	-271.151	5	5	80	A - 4.53	D - 4.20	G - 3.04	19	23	60.87	S - 2.71	A - 2.63	L - 2.09	59	68	30.88	Y - 8.27	A - 7.57	S - 1.21
254.1138	254.1178	0	1	0	A - 0.00	A - 0.00	A - 0.00	0	4	0	A - 0.00	A - 0.00	A - 0.00	2	108	0	E - 7.05	G - 5.69	S - 1.48
371.1788	371.1828	0	24	0	A - 0.00	A - 0.00	A - 0.00	53	145	22.07	R - 5.02	K - 3.81	A - 2.33	0	0	0	A - 0.00	A - 0.00	A - 0.00
212.1608	212.1648	31	31	100	G - 10.54	M - 3.07	N - 1.89	16	31	45.16	V - 3.38	G - 2.59	M - 2.04	0	0	0	A - 0.00	A - 0.00	A - 0.00
-0.0708	-0.0668	3	286	0	P - 7.88	E - 4.70	T - 3.03	23	1490	0.13	I - 3.01	V - 2.91	M - 1.99	4	378	0	V - 8.34	L - 2.53	A - 1.79
272.1464	272.1504	4	10	40	Y - 38.29	A - 0.00	A - 0.00	409	440	89.55	Y - 31.22	N - 0.71	D - 0.42	6	8	75	G - 11.81	Y - 2.13	A - 1.59
0.0882	0.0922	2	178	0.56	C - 7.45	L - 6.74	Q - 3.48	66	1037	0.58	C - 2.20	M - 2.07	V - 1.85	2	109	0.92	M - 22.92	V - 8.34	A - 0.00
-285.134	-285.13	11	13	46.15	D - 9.54	V - 3.03	G - 1.38	42	49	59.18	D - 5.70	G - 2.39	V - 2.08	37	38	81.58	P - 4.24	D - 4.07	S - 3.03
-298.166	-298.162	3	9	22.22	W - 8.29	Q - 6.96	D - 4.33	11	18	38.89	I - 4.06	W - 3.65	A - 3.18	27	59	33.9	Q - 5.55	A - 5.12	V - 4.53
-330.175	-330.171	1	2	0	Y - 9.57	F - 7.01	T - 4.55	26	36	19.44	M - 22.92	K - 4.79	A - 2.11	69	74	8.11	M - 28.08	K - 4.35	A - 1.74
421.0334	421.0374	0	0	0	A - 0.00	A - 0.00	A - 0.00	0	665	0	A - 0.00	A - 0.00	A - 0.00	0	0	0	A - 0.00	A - 0.00	A - 0.00
-0.0596	-0.0556	6	299	0.33	K - 2.93	V - 2.78	G - 2.53	37	1616	0.37	W - 4.34	V - 2.71	I - 2.38	1	463	0	I - 23.40	A - 0.00	A - 0.00
103.9226	103.9266	0	0	0	A - 0.00	A - 0.00	A - 0.00	3	8	0	D - 2.27	E - 2.22	T - 2.19	25	95	0	S - 4.89	E - 3.26	D - 3.17
0.1262	0.1302	0	112	0	A - 0.00	A - 0.00	A - 0.00	20	523	0.19	N - 4.21	M - 2.55	I - 2.51	0	49	0	A - 0.00	A - 0.00	A - 0.00
-216.077	-216.073	8	8	100	D - 9.18	T - 7.96	Y - 1.60	73	93	60.22	E - 4.05	D - 3.50	S - 3.21	56	59	79.66	E - 5.81	S - 3.88	A - 1.34
-313.203	-313.199	5	5	100	I - 7.80	S - 3.16	L - 2.70	133	146	65.75	I - 4.45	L - 4.02	T - 2.44	60	61	63.93	I - 11.63	L - 3.34	S - 1.91
166.0068	166.0108	30	34	50	Y - 10.81	Q - 5.37	K - 1.73	64	88	40.91	Y - 9.69	N - 5.19	E - 2.83	0	0	0	A - 0.00	A - 0.00	A - 0.00
121.975	121.979	4	12	33.33	T - 6.07	E - 4.70	S - 3.95	227	344	63.08	G - 3.67	S - 3.65	A - 2.79	4	7	28.57	T - 3.94	E - 3.05	S - 2.57
-442.244	-442.24	0	0	0	A - 0.00	A - 0.00	A - 0.00	11	17	47.06	T - 3.59	L - 2.91	E - 1.71	71	73	45.21	L - 6.29	V - 3.96	T - 0.85
-300.182	-300.178	3	5	20	T - 9.10	A - 4.77	N - 4.71	33	49	12.24	A - 8.71	K - 2.83	Q - 0.92	49	55	5.45	K - 8.52	A - 3.72	T - 3.53
70.0406	70.0446	7	16	37.5	A - 5.11	L - 2.89	S - 1.69	176	229	40.17	H - 12.87	D - 1.76	G - 1.53	8	11	9.09	D - 10.50	G - 6.64	A - 0.89
-486.246	-486.242	6	10	10	E - 7.59	S - 2.13	N - 1.54	47	94	9.57	E - 7.34	T - 2.94	S - 1.08	37	49	0	S - 6.41	E - 5.71	T - 0.49
-296.281	-296.277	0	0	0	A - 0.00	A - 0.00	A - 0.00	5	5	0	L - 10.11	A - 0.00	A - 0.00	77	77	0	L - 9.85	D - 0.27	P - 0.14
-475.28	-475.276	7	7	14.29	F - 20.03	L - 1.93	K - 0.84	42	45	44.44	F - 11.91	L - 2.83	V - 1.55	57	57	31.58	F - 20.17	V - 3.22	L - 0.89
-1.8794	-1.8754	2	64	0	Y - 19.14	V - 8.34	A - 0.00	14	313	2.24	D - 3.97	E - 1.98	G - 1.63	0	6	0	A - 0.00	A - 0.00	A - 0.00
29.0128	29.0168	11	19	5.26	G - 4.22	C - 4.06	Y - 2.61	144	260	10.77	N - 3.90	A - 2.98	C - 2.51	41	42	0	N - 11.37	A - 8.54	A - 0.00
242.1802	242.1842	1	8	12.5	D - 10.50	Q - 5.22	T - 4.55	350	660	26.06	W - 4.24	E - 2.59	D - 2.09	0	1	0	A - 0.00	A - 0.00	A - 0.00
181.038	181.042	0	32	0	A - 0.00	A - 0.00	A - 0.00	1	7	14.29	F - 5.61	D - 4.20	A - 2.86	0	0	0	A - 0.00	A - 0.00	A - 0.00
71.983	71.987	34	34	97.06	W - 61.05	N - 1.66	T - 0.62	14	23	39.13	W - 37.30	Q - 4.48	S - 1.61	1	5	20	K - 8.79	A - 7.15	A - 0.00
-357.156	-357.152	4	4	100	E - 12.33	L - 1.26	A - 0.00	28	39	61.54	E - 6.33	L - 2.71	I - 2.09	66	68	75	E - 5.38	L - 3.08	D - 1.63
249.9566	249.9606	0	1	0	A - 0.00	A - 0.00	A - 0.00	1	70	0	H - 6.39	E - 4.70	A - 2.38	7	77	0	E - 4.03	A - 3.00	L - 2.89
-443.204	-443.2	2	2	100	T - 7.59	D - 5.25	N - 4.71	43	105	29.52	N - 5.36	T - 2.17	A - 1.99	40	63	41.27	E - 3.90	D - 2.80	S - 2.55
3.0322	3.0362	10	81	1.23	C - 22.36	I - 2.34	N - 1.41	401	1509	0.4	C - 2.07	F - 1.49	N - 1.38	22	32	62.5	C - 23.37	V - 6.45	F - 2.55
319.0096	319.0136	0	50	0	A - 0.00	A - 0.00	A - 0.00	0	1	0	A - 0.00	A - 0.00	A - 0.00	0	0	0	A - 0.00	A - 0.00	A - 0.00
-0.1758	-0.1718	0	45	0	A - 0.00	A - 0.00	A - 0.00	6	249	0	E - 11.74	H - 1.06	D - 0.58	0	119	0	A - 0.00	A - 0.00	A - 0.00
1.0772	1.0812	3	187	0	S - 3.53	D - 3.50	A - 3.06	136	1218	0.82	I - 2.97	Y - 1.77	E - 1.65	2	169	0	I - 17.55	L - 2.53	A - 0.00
-471.235	-471.231	0	1	0	A - 0.00	A - 0.00	A - 0.00	16	27	14.81	D - 8.20	L - 2.90	A - 2.31	66	71	45.07	D - 2.93	G - 2.45	A - 2.09
126.8968	126.9008	28	39	5.13	Y - 27.05	W - 2.87	D - 0.94	8	14	14.29	Y - 23.13	E - 2.64	D - 1.31	0	2	0	A - 0.00	A - 0.00	A - 0.00
55.9214	55.9254	23	118	0	N - 3.35	C - 2.38	D - 2.04	114	285	0.7	D - 2.26	E - 2.23	A - 1.76	3	4	0	E - 4.70	Y - 4.25	Q - 2.32
-0.1202	-0.1162	1	125	0	P - 7.88	N - 7.06	I - 5.85	12	637	0.31	H - 3.19	L - 2.74	Q - 2.61	8	265	0	F - 12.27	E - 4.40	I - 2.92
-3.9746	-3.9706	0	51	0	A - 0.00	A - 0.00	A - 0.00	15	282	0.35	H - 6.47	I - 3.43	V - 2.41	1	73	0	N - 28.25	A - 0.00	A - 0.00

288.1548	288.1588	0	1	0	A - 0.00	A - 0.00	A - 0.00	11	24	29.17	Y - 2.44	Q - 1.60	D - 1.57	0	106	0	A - 0.00	A - 0.00	A - 0.00
27.0452	27.0492	1	4	0	T - 9.10	A - 7.15	A - 0.00	561	600	7	C - 9.17	T - 8.06	H - 2.82	2	5	0	A - 10.72	F - 7.01	A - 0.00
-341.198	-341.194	4	6	50	L - 5.90	I - 2.92	D - 2.62	51	62	67.74	L - 5.37	I - 4.39	V - 1.87	55	58	93.1	L - 5.88	V - 2.62	I - 2.08
2.087	2.091	0	130	0	A - 0.00	A - 0.00	A - 0.00	126	816	0.12	D - 1.99	I - 1.42	Y - 1.38	4	48	2.08	Y - 9.57	N - 4.94	F - 3.51
-301.166	-301.162	3	3	100	T - 10.11	I - 3.90	D - 3.50	23	32	37.5	S - 4.43	T - 4.42	E - 1.84	55	66	57.58	T - 8.42	I - 5.44	S - 1.50
73.0132	73.0172	3	13	15.38	M - 15.28	F - 6.23	P - 5.25	67	89	7.87	A - 10.17	F - 1.78	Y - 1.19	20	89	0	M - 7.03	G - 6.22	H - 2.04
-336.194	-336.19	6	8	75	Y - 18.08	R - 8.33	C - 2.48	102	108	81.48	Y - 12.83	C - 6.43	R - 5.26	37	38	76.32	Y - 12.42	R - 5.72	H - 4.83
-26.0192	-26.0152	15	21	0	P - 6.98	A - 4.61	T - 4.27	148	181	0	A - 7.00	T - 3.97	P - 3.53	25	25	0	P - 15.44	D - 0.42	A - 0.00
140.0572	140.0612	3	3	100	L - 10.11	A - 0.00	A - 0.00	13	23	52.17	L - 8.75	V - 1.28	T - 0.35	46	53	86.79	L - 10.11	A - 0.00	A - 0.00
-340.225	-340.221	3	7	28.57	G - 5.06	H - 4.26	I - 3.90	81	101	23.76	R - 6.05	L - 2.89	I - 1.21	50	64	3.13	R - 8.12	L - 2.87	A - 2.00
300.1418	300.1458	2	16	12.5	Y - 19.14	L - 5.06	A - 0.00	133	284	38.03	T - 4.75	D - 3.12	F - 1.93	1	2	50	E - 4.70	S - 3.95	L - 3.37
-331.159	-331.155	4	5	20	K - 4.39	M - 3.82	A - 2.09	7	15	13.33	M - 30.56	I - 4.46	A - 0.68	65	67	20.9	M - 36.05	I - 1.02	S - 0.55
-1.9442	-1.9402	5	116	0.86	F - 5.61	I - 4.68	S - 2.14	57	833	1.8	Q - 3.06	I - 2.63	D - 2.45	0	39	0	A - 0.00	A - 0.00	A - 0.00
-3.996	-3.992	17	48	14.58	T - 3.48	L - 1.98	G - 1.69	37	269	7.81	V - 4.11	T - 3.81	D - 3.57	7	61	6.56	T - 6.50	F - 6.01	P - 3.38
-0.0846	-0.0806	3	223	0.45	L - 5.62	I - 3.90	K - 2.93	20	1270	0.16	V - 2.95	Q - 2.50	T - 2.48	2	363	0	L - 5.06	T - 4.55	E - 3.52
-48.1294	-48.1254	10	13	0	S - 4.90	V - 2.78	L - 2.29	51	120	0.83	S - 4.35	P - 2.41	K - 1.76	13	40	0	S - 4.87	T - 4.20	Q - 3.21
-257.103	-257.099	8	8	50	N - 7.06	E - 3.52	Y - 2.39	108	116	69.83	G - 5.25	E - 4.94	N - 4.45	41	41	97.56	E - 6.80	G - 3.70	Q - 3.31
-327.194	-327.19	5	6	83.33	Q - 10.45	A - 7.15	A - 0.00	19	33	30.3	A - 5.66	Q - 4.31	K - 2.70	54	55	80	Q - 8.66	A - 6.67	K - 1.30
-269.187	-269.183	2	4	50	R - 8.82	I - 5.85	L - 2.53	121	177	10.17	R - 9.30	I - 3.31	L - 2.64	38	43	4.65	R - 14.51	L - 1.53	I - 0.46
-372.203	-372.199	13	14	14.29	T - 9.31	A - 3.30	I - 1.35	28	37	32.43	T - 6.94	A - 3.70	S - 1.21	22	27	11.11	T - 14.37	I - 1.23	S - 0.85
-326.197	-326.193	6	8	0	T - 6.07	K - 5.86	N - 4.71	43	57	17.54	V - 5.17	T - 3.86	K - 2.50	50	59	62.71	N - 10.92	V - 4.23	L - 2.56
384.2106	384.2146	4	4	100	E - 7.05	D - 2.62	T - 2.28	244	266	85.34	E - 4.75	C - 3.55	D - 2.82	26	30	70	S - 5.93	E - 5.15	Y - 1.10
1.8654	1.8694	2	91	1.1	V - 8.34	L - 5.06	A - 0.00	11	445	0.22	V - 3.54	L - 2.91	E - 2.39	2	123	0.81	A - 7.15	L - 5.06	A - 0.00
-332.187	-332.183	4	5	0	F - 17.53	G - 5.69	A - 0.00	15	33	9.09	F - 10.66	G - 3.61	K - 1.17	43	50	2	F - 19.13	K - 4.84	G - 0.65
0.8922	0.8962	1	166	0	Q - 4.18	P - 3.15	G - 3.04	21	944	0	V - 4.43	I - 2.72	M - 2.18	6	291	0	Y - 6.38	V - 5.56	F - 4.67
171.1238	171.1278	0	2	0	A - 0.00	A - 0.00	A - 0.00	10	25	24	Y - 3.89	G - 3.05	N - 2.88	1	81	0	T - 6.07	P - 5.25	E - 4.70
112.0962	112.1002	8	10	10	M - 18.84	S - 1.99	C - 1.92	485	529	40.45	M - 18.41	E - 1.24	D - 1.20	38	44	38.64	M - 13.85	Y - 2.55	E - 1.51
226.1192	226.1232	0	0	0	A - 0.00	A - 0.00	A - 0.00	2	34	0	I - 3.90	Y - 3.83	L - 3.71	0	88	0	A - 0.00	A - 0.00	A - 0.00
44.9832	44.9872	8	28	3.57	W - 22.60	Y - 8.77	I - 2.19	306	421	12.11	Y - 25.67	W - 2.05	N - 0.80	55	62	4.84	Y - 30.98	M - 1.67	R - 1.05
61.9666	61.9706	0	2	0	A - 0.00	A - 0.00	A - 0.00	4	9	11.11	W - 6.70	R - 4.41	L - 4.21	58	69	0	R - 15.92	L - 0.76	D - 0.36
-1.9604	-1.9564	4	126	0.79	L - 5.06	P - 3.94	T - 1.52	34	904	1.33	W - 4.23	M - 3.54	Y - 2.84	1	60	0	M - 22.92	A - 7.15	A - 0.00
29.9978	30.0018	31	64	1.56	S - 3.44	T - 2.04	A - 2.00	349	654	4.59	S - 2.39	T - 1.65	D - 1.24	0	2	0	A - 0.00	A - 0.00	A - 0.00
17.9802	17.9842	22	40	7.5	N - 4.87	M - 2.66	A - 1.83	80	226	7.52	M - 3.71	W - 2.60	N - 2.32	1	6	0	M - 9.17	N - 5.65	S - 4.75
4.0072	4.0112	9	70	0	C - 4.97	V - 3.09	G - 1.83	476	2291	0.17	Q - 1.96	I - 1.53	M - 1.46	0	1	0	A - 0.00	A - 0.00	A - 0.00
0.903	0.907	0	182	0	A - 0.00	A - 0.00	A - 0.00	28	1066	0.09	L - 2.87	I - 2.42	Q - 2.24	3	348	0	V - 11.12	I - 7.80	A - 0.00
40.0288	40.0328	5	10	50	N - 14.13	L - 5.06	A - 0.00	324	338	93.2	G - 3.06	N - 1.86	L - 1.74	4	4	100	A - 7.15	G - 3.16	E - 2.94
-33.0236	-33.0196	8	14	0	W - 16.74	M - 11.98	N - 6.21	129	150	54.67	M - 11.69	Q - 3.97	C - 2.21	25	27	44.44	M - 13.10	Q - 3.69	W - 3.21
-72.9994	-72.9954	9	9	22.22	C - 36.43	N - 1.05	V - 0.93	99	186	4.84	C - 27.37	M - 2.24	E - 1.83	37	63	0	C - 36.50	S - 1.67	T - 0.34
-301.147	-301.143	0	1	0	A - 0.00	A - 0.00	A - 0.00	8	17	5.88	L - 7.03	T - 3.55	I - 0.52	71	72	95.83	M - 14.76	V - 6.51	A - 3.46
115.0442	115.0482	4	12	0	A - 6.55	K - 3.66	S - 2.97	27	114	0.88	K - 2.66	I - 2.46	Y - 2.11	3	40	5	A - 6.67	E - 4.70	C - 2.98
-473.214	-473.21	3	3	0	A - 11.91	T - 3.03	A - 0.00	53	81	19.75	T - 5.94	A - 5.30	E - 2.64	49	51	31.37	A - 6.08	T - 3.84	E - 1.44
-370.187	-370.183	2	2	50	Q - 10.45	E - 3.52	L - 2.53	37	42	71.43	V - 5.75	Q - 2.73	G - 2.22	55	71	42.25	Q - 8.15	I - 7.59	E - 1.67
-0.9288	-0.9248	6	185	1.08	M - 15.28	Q - 4.06	I - 3.90	45	1084	1.29	V - 2.62	M - 2.39	Y - 2.30	3	117	0	F - 9.35	D - 3.50	P - 2.63
294.1012	294.1052	0	6	0	A - 0.00	A - 0.00	A - 0.00	0	36	0	A - 0.00	A - 0.00	A - 0.00	0	43	0	A - 0.00	A - 0.00	A - 0.00
270.1436	270.1476	2	4	25	E - 10.57	V - 4.17	A - 0.00	358	372	90.86	E - 2.80	T - 2.21	V - 2.20	19	58	29.31	W - 9.87	G - 4.79	Y - 4.03
383.2264	383.2304	6	7	42.86	T - 5.31	D - 5.25	A - 2.38	295	307	92.18	A - 4.81	D - 4.78	V - 1.34	14	15	93.33	D - 9.00	W - 5.74	C - 3.19
322.1612	322.1652	0	1	0	A - 0.00	A - 0.00	A - 0.00	1	16	6.25	R - 17.65	A - 0.00	A - 0.00	0	71	0	A - 0.00	A - 0.00	A - 0.00
-441.235	-441.231	6	8	75	N - 9.42	Q - 6.96	A - 4.77	29	37	37.84	A - 2.91	D - 2.82	N - 2.54	41	41	68.29	Q - 9.85	N - 6.43	A - 3.25
-276.149	-276.145	14	14	100	Y - 18.23	I - 3.62	L - 3.25	40	47	80.85	Y - 17.87	L - 4.59	I - 1.07	19	23	69.57	Y - 17.47	L - 3.90	I - 2.46
60.0008	60.0048	4	18	5.56	W - 5.02	M - 4.20	H - 3.51	15	172	2.33	M - 10.88	H - 1.92	S - 1.79	3	14	0	A - 9.53	Q - 3.48	K - 2.93
355.1092	355.1132	0	0	0	A - 0.00	A - 0.00	A - 0.00	7	10	70	L - 7.77	Q - 1.49	A - 1.02	63	63	100	L - 10.11	A - 0.00	A - 0.00
82.9232	82.9272	0	0	0	A - 0.00	A - 0.00	A - 0.00	1	5	0	L - 6.74	A - 4.77	A - 0.00	71	95	0	G - 3.40	I - 3.15	A - 2.76
1.829	1.833	1	51	0	L - 10.11	A - 0.00	A - 0.00	3	227	0	V - 5.56	A - 4.77	I - 1.56	0	70	0	A - 0.00	A - 0.00	A - 0.00
202.0756	202.0796	1	7	0	Q - 10.45	T - 9.10	A - 0.00	9	194	2.58	R - 5.88	F - 4.01	E - 4.00	4	43	6.98	Q - 15.67	I - 5.85	A - 0.00
-45.0234	-45.0194	2	2	100	D - 6.12	H - 4.79	F - 4.67	338	344	95.35	C - 20.07	D - 1.79	I - 1.58	39	40	90	D - 6.77	Q - 3.39	A - 2.57
-0.8846	-0.8806	0	109	0	A - 0.00	A - 0.00	A - 0.00	19	488	0.82	Y - 4.32	F - 3.50	M - 3.02	0	29	0	A - 0.00	A - 0.00	A - 0.00
120.102	120.106	0	0	0	A - 0.00	A - 0.00	A - 0.00	0	4	0	A - 0.00	A - 0.00	A - 0.00	0	68	0	A - 0.00	A - 0.00	A - 0.00
39.9918	39.9958	12	22	50	N - 10.99	H - 3.19	C - 1.86	24	35	54.29	N - 4.37	H - 3.77	F - 3.57	3	35	2.86	S - 3.95	D - 3.50	C - 2.48
-344.172	-344.168	6	7	14.29	K - 7.32	E - 4.70	T - 3.03	56	117	20.51	K - 6.22	E - 4.05	S - 2.57	39	43	27.91	K - 7.62	D - 4.93	E - 2.32
346.038	346.042	0	0	0	A - 0.00	A - 0.00	A - 0.00	1	578	0	A - 5.72	F - 5.61	D - 4.20	0	1	0	A - 0.00	A - 0.00	A - 0.00
-147.037	-147.033	0	2	0	A - 0.00	A - 0.00	A - 0.00	39	110	27.27	M - 21.35	S - 1.72	Y - 1.55	51	57	84.21	M - 24.53	L - 2.18	A - 0.92

Table C-2 List of top 500 detected features in mass shift histogram with potential explanations.

Peak Apex	Peptides			Decoy Peptides			FDR (%)					
	HEK293	TNBC	HeLa	HEK293	TNBC	HeLa	HEK293	TNBC	HeLa	Potential Modification 1	Potential Modification 2	Potential Modification 3
-0.0002	88477	166767	60860	94	190	63	0.11%	0.11%	0.10%			
1.0026	18709	76569	2325	0	18	0	0.00%	0.02%	0.00%	First isotopic peak		
43.0058	7288	16447	9	3	6	0	0.04%	0.04%	0.00%	Carbamylation/Ala->Asn substitution		
0.984	1789	9633	1870	2	17	2	0.11%	0.18%	0.11%	Deamidation/Asn->Asp substitution/Gln->Glu substitution		
-17.0268	2384	4721	1138	1	2	1	0.04%	0.04%	0.09%	Pyro-glu from Q/Loss of ammonia		
52.9128	1	1567	3221	0	0	0	0.00%	0.00%	0.00%	Unannotated mass-shift 52.9128		
-128.095	627	3185	1172	2	15	12	0.32%	0.47%	1.02%	Loss of C-terminal K from Heavy Chain of MAb/Deletion of K		
2.0052	3916	31949	164	0	8	0	0.00%	0.03%	0.00%	Second isotopic peak		
128.0948	707	4104	601	0	9	0	0.00%	0.22%	0.00%	Addition of lysine due to transpeptidation/Addition of K		
15.9948	978	1490	1314	1	5	1	0.10%	0.34%	0.08%	Oxidation or Hydroxylation/Ala->Ser substitution/Phe->Tyr substitution		
31.99	747	619	1114	1	1	1	0.13%	0.16%	0.09%	dihydroxy/Pro->Glu substitution		
301.9864	2100	681	2	0	1	0	0.00%	0.15%	0.00%	Unannotated mass-shift 301.9864		
27.995	1537	2856	65	8	6	3	0.52%	0.21%	4.62%	Formylation/Ser->Asp substitution/Thr->Glu substitution		
79.9664	644	1350	642	1	4	3	0.16%	0.30%	0.47%	Phosphorylation		
53.9186	1549	607	143	0	0	0	0.00%	0.00%	0.00%	Replacement of 2 protons by iron		
183.0352	1242	1433	25	0	0	0	0.00%	0.00%	0.00%	Aminoethylbenzenesulfonylation		
57.0214	538	156	1135	0	1	1	0.00%	0.64%	0.09%	Iodoacetamide derivative/Ala->Gln substitution/Gly->Asn substitution/Addition of Glycine/Addition of G		
23.958	0	52	1098	0	1	0	0.00%	1.92%	0.00%	Unannotated mass-shift 23.9580		
-18.0104	325	803	728	3	2	1	0.92%	0.25%	0.14%	Dehydration/Pyro-glu from E		
156.1012	304	1437	215	0	1	0	0.00%	0.07%	0.00%	Addition of arginine due to transpeptidation/Addition of R		
284.1268	0	194	951	0	0	0	0.00%	0.00%	0.00%	Unannotated mass-shift 284.1268		
162.1258	0	2	600	0	0	1	0.00%	0.00%	0.17%	Unannotated mass-shift 162.1258		
17.0258	125	758	445	2	8	2	1.60%	1.06%	0.45%	replacement of proton with ammonium ion		
234.0742	0	4	495	0	0	0	0.00%	0.00%	0.00%	Unannotated mass-shift 234.0742		
37.9472	129	1012	659	1	0	0	0.78%	0.00%	0.00%	Replacement of 2 protons by calcium		
306.0952	10	5	766	0	0	0	0.00%	0.00%	0.00%	Unannotated mass-shift 306.0952		
44.0084	861	4342	1	0	1	0	0.00%	0.02%	0.00%	S-Ethylcystine from Serine/Ser->Met substitution		
-15.9956	138	1910	285	0	6	0	0.00%	0.31%	0.00%	reduction/Ser->Ala substitution/Tyr->Phe substitution		
302.9896	763	112	0	0	1	0	0.00%	0.89%	0.00%	First isotopic peak	Unannotated mass-shift 301.9864	
-71.0368	150	412	67	2	1	0	1.33%	0.24%	0.00%	Gln->Gly substitution/Deletion of A		
249.981	475	21	1	0	0	0	0.00%	0.00%	0.00%	Unannotated mass-shift 249.9810		
-113.0836	231	630	79	2	1	0	0.87%	0.16%	0.00%	Deletion of I/Deletion of L		
21.9804	178	1599	83	0	1	0	0.00%	0.06%	0.00%	Sodium adduct		
241.1788	35	455	15	0	1	0	0.00%	0.22%	0.00%	Unannotated mass-shift 241.1788		
-145.1218	122	207	107	0	0	0	0.00%	0.00%	0.00%	Pyro-glu from Q/Loss of ammonia	Loss of C-terminal K from Heavy Chain of MAb/Deletion of K	
14.0156	84	385	55	1	1	0	1.19%	0.26%	0.00%	Methylation/Asp->Glu substitution/Gly->Ala substitution/Ser->Thr substitution/Val->Leu/Ile substitution/Asn->Gln substitution		
1.9882	717	6523	107	0	8	0	0.00%	0.12%	0.00%	First isotopic peak	Deamidation/Asn->Asp substitution/Gln->Glu substitution	
-2.016	266	1324	137	1	5	0	0.38%	0.38%	0.00%	2-amino-3-oxo-butanoic_acid/Val->Pro substitution		
151.9956	4	1847	2	0	0	0	0.00%	0.00%	0.00%	Unannotated mass-shift 151.9956		
3.0076	591	8018	18	0	5	0	0.00%	0.06%	0.00%	Unannotated mass-shift 3.0076		
-85.0894	242	524	0	2	2	1	0.83%	0.38%	100.00%	Carbamylation/Ala->Asn substitution	Loss of C-terminal K from Heavy Chain of MAb/Deletion of K	
-16.0246	367	1744	30	0	1	0	0.00%	0.06%	0.00%	First isotopic peak	Pyro-glu from Q/Loss of ammonia	
16.9968	396	776	29	1	2	0	0.25%	0.26%	0.00%	Asn->Met substitution		

-1.0036	481	2749	530	4	16	4	0.83%	0.58%	0.76%	Pyro-glu from Q/Loss of ammonia	Second isotopic peak	Methylation/Asp->Glu substitution/Gly->Ala substitution/Ser->Thr substitution/Val->Leu/Ile substitution/Asn->Gln substitution
54.9218	456	259	14	0	0	0	0.00%	0.00%	0.00%	First isotopic peak	Replacement of 2 protons by iron	
80.9702	290	821	4	0	1	0	0.00%	0.12%	0.00%	First isotopic peak	Phosphorylation	
204.136	0	4	233	0	0	1	0.00%	0.00%	0.43%	Acetylation/Ser->Glu substitution	Unannotated mass-shift 162.1258	
131.0396	5	186	58	0	0	0	0.00%	0.00%	0.00%	Addition of M		
125.8968	249	43	1	0	0	0	0.00%	0.00%	0.00%	Iodination		
184.037	337	684	2	0	0	0	0.00%	0.00%	0.00%	First isotopic peak	Aminoethylbenzenesulfonylation	
129.0974	134	1650	3	0	1	0	0.00%	0.06%	0.00%	First isotopic peak	Addition of lysine due to transpeptidation/Addition of K	
216.1002	1	10	202	0	0	0	0.00%	0.00%	0.00%	Unannotated mass-shift 216.1002		
-0.9836	471	2073	153	2	18	4	0.43%	0.87%	2.61%	Amidation/Asp->Asn substitution/Glu->Gln substitution		
197.0454	2	13	5	0	1	0	0.00%	7.69%	0.00%	glycerylphosphorylethanolamine		
3.9946	151	755	83	0	1	1	0.00%	0.13%	1.21%	tryptophan oxidation to kynurenin/Pro->Thr substitution		
203.08	94	178	59	1	0	0	1.06%	0.00%	0.00%	N-Acetylhexosamine		
-14.0156	51	450	46	1	1	0	1.96%	0.22%	0.00%	Ala->Gly substitution/Glu->Asp substitution/Leu/Ile->Val substitution/Thr->Ser substitution/Gln->Asn substitution		
-1.032	386	1702	214	0	7	0	0.00%	0.41%	0.00%	Lysine oxidation to aminoadipic semialdehyde		
-99.069	125	257	20	0	1	0	0.00%	0.39%	0.00%	Deletion of V		
41.0264	56	63	225	0	0	0	0.00%	0.00%	0.00%	amidination of lysines or N-terminal amines with methyl acetimidate/Ser->Gln substitution		
173.0506	27	230	70	0	1	0	0.00%	0.44%	0.00%	Acetylation/Ser->Glu substitution	Addition of M	
0.9482	310	1168	414	1	14	2	0.32%	1.20%	0.48%	Lys->Glu substitution		
28.0314	58	225	13	0	4	0	0.00%	1.78%	0.00%	di-Methylation/Acetaldehyde +28/Ethylation/Ala->Val substitution/Cys->Met substitution		
248.1262	4	7	182	0	0	0	0.00%	0.00%	0.00%	Unannotated mass-shift 216.1002	tryptophan oxidation to kynurenin/Pro->Thr substitution	di-Methylation/Acetaldehyde +28/Ethylation/Ala->Val substitution/Cys->Met substitution
28.9974	288	991	7	0	0	0	0.00%	0.00%	0.00%	First isotopic peak	Formylation/Ser->Asp substitution/Thr->Glu substitution	
-328.211	15	61	37	0	0	0	0.00%	0.00%	0.00%	Unannotated mass-shift -328.2110		
-156.1012	34	500	96	1	3	0	2.94%	0.60%	0.00%	Loss of arginine due to transpeptidation/Deletion of R		
-229.143	10	63	51	0	1	0	0.00%	1.59%	0.00%	Unannotated mass-shift -229.1430		
-127.1114	33	146	54	0	0	0	0.00%	0.00%	0.00%	Deamidation/Asn->Asp substitution/Gln->Glu substitution	Loss of C-terminal K from Heavy Chain of MAb/Deletion of K	
11.9996	62	356	71	0	0	0	0.00%	0.00%	0.00%	Addition of Carbon to cysteine		
43.99	137	552	0	1	1	0	0.73%	0.18%	0.00%	Carboxylation/Ala->Asp substitution		
-170.1056	80	92	24	0	0	0	0.00%	0.00%	0.00%	Gln->Gly substitution/Deletion of A	Deletion of V	
284.1964	31	155	13	0	0	0	0.00%	0.00%	0.00%	Addition of lysine due to transpeptidation/Addition of K	Addition of arginine due to transpeptidation/Addition of R	
-257.1376	27	110	58	0	0	1	0.00%	0.00%	1.72%	Unannotated mass-shift -257.1376		
32.9924	215	258	22	0	0	0	0.00%	0.00%	0.00%	First isotopic peak	dihydroxy/Pro->Glu substitution	
-57.0216	24	536	26	0	0	0	0.00%	0.00%	0.00%	Asn->Gly substitution/Gln->Ala substitution/Deletion of G		
-0.0362	273	1575	360	4	10	0	1.47%	0.64%	0.00%	Lys->Gln substitution		
171.101	104	399	3	1	0	0	0.96%	0.00%	0.00%	Carbamylation/Ala->Asn substitution	Addition of lysine due to transpeptidation/Addition of K	
-199.1322	19	133	85	0	0	0	0.00%	0.00%	0.00%	Loss of C-terminal K from Heavy Chain of MAb/Deletion of K	Gln->Gly substitution/Deletion of A	
-129.0424	21	100	92	0	4	0	0.00%	4.00%	0.00%	Deletion of E		
0.0366	266	4260	223	1	6	0	0.38%	0.14%	0.00%	Gln->Lys substitution		
378.1146	3	3	244	0	1	0	0.00%	33.33%	0.00%	Unannotated mass-shift 306.0952	di-Methylation/Acetaldehyde +28/Ethylation/Ala->Val substitution/Cys->Met substitution	Carboxylation/Ala->Asp substitution

-9.037	95	251	17	1	0	0	1.05%	0.00%	0.00%	Deamidation/Asn->Asp substitution/Gln->Glu substitution	tryptophan oxidation to kynurenin/Pro->Thr substitution	Ala->Gly substitution/Glu->Asp substitution/Leu/Ile->Val substitution/Thr->Ser substitution/Gln->Asn substitution
-87.032	31	127	45	1	1	0	3.23%	0.79%	0.00%	Deletion of S		
-184.1208	62	78	46	0	0	0	0.00%	0.00%	0.00%	Gln->Gly substitution/Deletion of A	Deletion of I/Deletion of L	
-200.116	60	81	26	0	1	0	0.00%	1.24%	0.00%	Addition of lysine due to transpeptidation/Addition of K	Unannotated mass-shift -328.2110	
170.0948	1	51	150	0	0	0	0.00%	0.00%	0.00%	Unannotated mass-shift 216.1002	amidination of lysines or N-terminal amines with methyl acetimidate/Ser->Gln substitution	Deletion of S
244.131	4	82	186	0	0	0	0.00%	0.00%	0.00%	Unannotated mass-shift 216.1002	di-Methylation/Acetaldehyde +28/Ethylation/Ala->Val substitution/Cys->Met substitution	
113.0836	32	194	13	4	3	0	12.50%	1.55%	0.00%	Acetylhypusine/Addition of I/Addition of L		
-114.0424	31	115	54	2	0	0	6.45%	0.00%	0.00%	Deletion of N		
260.127	1	12	176	0	0	0	0.00%	0.00%	0.00%	Oxidation or Hydroxylation/Ala->Ser substitution/Phe->Tyr substitution	Unannotated mass-shift 216.1002	di-Methylation/Acetaldehyde +28/Ethylation/Ala->Val substitution/Cys->Met substitution
68.9078	0	19	115	0	0	0	0.00%	0.00%	0.00%	Unannotated mass-shift 52.9128	Oxidation or Hydroxylation/Ala->Ser substitution/Phe->Tyr substitution	
157.1042	84	711	0	0	1	0	0.00%	0.14%	0.00%	First isotopic peak	Addition of arginine due to transpeptidation/Addition of R	
269.1854	21	180	29	0	1	0	0.00%	0.56%	0.00%	Addition of arginine due to transpeptidation/Addition of R	Acetylhypusine/Addition of I/Addition of L	
-112.1004	20	62	61	1	0	0	5.00%	0.00%	0.00%	Deamidation/Asn->Asp substitution/Gln->Glu substitution	Deletion of I/Deletion of L	
-128.0588	43	147	39	1	4	0	2.33%	2.72%	0.00%	Deletion of Q		
42.0104	49	279	37	0	1	0	0.00%	0.36%	0.00%	Acetylation/Ser->Glu substitution		
-215.1272	7	97	86	0	0	0	0.00%	0.00%	0.00%	Acetylation/Ser->Glu substitution	Unannotated mass-shift -257.1376	
-115.027	45	370	27	0	2	0	0.00%	0.54%	0.00%	Deletion of D		
50.0002	2	38	172	0	0	0	0.00%	0.00%	0.00%	Acetylation/Ser->Glu substitution	tryptophan oxidation to kynurenin/Pro->Thr substitution	tryptophan oxidation to kynurenin/Pro->Thr substitution
0.0474	287	2384	197	0	3	0	0.00%	0.13%	0.00%	replacement of proton with ammonium ion	reduction/Ser->Ala substitution/Tyr->Phe substitution	Amidation/Asp->Asn substitution/Glu->Gln substitution
0.9376	307	1089	317	2	3	0	0.65%	0.28%	0.00%	Deamidation/Asn->Asp substitution/Gln->Glu substitution	Deamidation/Asn->Asp substitution/Gln->Glu substitution	Lysine oxidation to aminoadipic semialdehyde
210.1616	0	1030	2	0	0	0	0.00%	0.00%	0.00%	Unannotated mass-shift 210.1616		
53.8972	1	73	52	0	0	0	0.00%	0.00%	0.00%	Deamidation/Asn->Asp substitution/Gln->Glu substitution	Unannotated mass-shift 52.9128	
308.0824	7	58	182	0	0	0	0.00%	0.00%	0.00%	First isotopic peak	Deamidation/Asn->Asp substitution/Gln->Glu substitution	Unannotated mass-shift 306.0952
13.979	39	130	156	0	1	0	0.00%	0.77%	0.00%	proline oxidation to pyroglutamic acid/Tryptophan oxidation to oxolactone/Thr->Asp substitution		
-241.1796	26	138	66	0	0	0	0.00%	0.00%	0.00%	Loss of C-terminal K from Heavy Chain of MAb/Deletion of K	Deletion of I/Deletion of L	
420.0506	2	1116	1	0	2	0	0.00%	0.18%	0.00%	Unannotated mass-shift 420.0506		
-163.0632	19	96	38	0	0	0	0.00%	0.00%	0.00%	Deletion of Y		
-286.1642	7	40	27	0	0	0	0.00%	0.00%	0.00%	Unannotated mass-shift -229.1430	Asn->Gly substitution/Gln->Ala substitution/Deletion of G	
2.9892	208	2511	14	0	6	0	0.00%	0.24%	0.00%	glycosylated asparagine 18O labeling/Deamidation in presence of O18		
215.127	12	180	12	0	0	0	0.00%	0.00%	0.00%	Unannotated mass-shift 215.1270		
-372.2198	1	19	12	0	0	0	0.00%	0.00%	0.00%	Unannotated mass-shift -372.2198		
-146.1054	31	45	61	4	5	38	12.90%	11.11%	62.30%	Loss of C-terminal K from Heavy Chain of MAb/Deletion of K	Dehydration/Pyro-glu from E	
-227.1642	16	54	27	1	1	0	6.25%	1.85%	0.00%	Loss of C-terminal K from Heavy Chain of MAb/Deletion of K	Deletion of V	
213.1002	1	56	173	0	2	0	0.00%	3.57%	0.00%	Dehydration/Pyro-glu from E	N-Acetylhexosamine	di-Methylation/Acetaldehyde +28/Ethylation/Ala->Val substitution/Cys->Met substitution

-243.1218	9	59	16	0	0	1	0.00%	0.00%	6.25%	Loss of C-terminal K from Heavy Chain of MAb/Deletion of K	Deletion of D	
250.9822	173	15	1	0	0	0	0.00%	0.00%	0.00%	Aminoethylbenzenesulfonylation	Unannotated mass-shift 23.9580	Carboxylation/Ala->Asp substitution
-212.1526	53	47	6	0	0	0	0.00%	0.00%	0.00%	Deletion of I/Deletion of L	Deletion of V	
-71.0732	46	13	51	0	0	1	0.00%	0.00%	1.96%	Lys->Gly substitution		
227.1638	16	185	5	0	0	0	0.00%	0.00%	0.00%	Unannotated mass-shift 241.1788	Ala->Gly substitution/Glu->Asp substitution/Leu/Ile->Val substitution/Thr->Ser substitution/Gln->Asn substitution	
-131.041	7	115	36	0	0	0	0.00%	0.00%	0.00%	Removal of initiator methionine from protein N-terminus/Deletion of M		
16.9782	50	189	41	0	3	2	0.00%	1.59%	4.88%	Deamidation/Asn->Asp substitution/Gln->Glu substitution	Oxidation or Hydroxylation/Ala->Ser substitution/Phe->Tyr substitution	
27.011	25	237	13	0	2	0	0.00%	0.84%	0.00%	Ser->Asn substitution/Thr->Gln substitution		
-91.0096	3	92	75	0	0	0	0.00%	0.00%	0.00%	di-Methylation/Acetaldehyde +28/Ethylation/Ala->Val substitution/Cys->Met substitution	Addition of Carbon to cysteine	Removal of initiator methionine from protein N-terminus/Deletion of M
0.9262	252	1015	332	0	6	0	0.00%	0.59%	0.00%	Lysine oxidation to aminoadipic semialdehyde	Lysine oxidation to aminoadipic semialdehyde	glycosylated asparagine 18O labeling/Deamidation in presence of O18
-101.0474	18	91	39	1	1	0	5.56%	1.10%	0.00%	Deletion of T		
-142.0746	16	49	39	0	0	0	0.00%	0.00%	0.00%	Gln->Gly substitution/Deletion of A	Gln->Gly substitution/Deletion of A	
-204.0902	14	25	11	0	0	0	0.00%	0.00%	0.00%	Oxidation or Hydroxylation/Ala->Ser substitution/Phe->Tyr substitution	Asn->Gly substitution/Gln->Ala substitution/Deletion of G	Deletion of Y
227.1382	7	109	25	0	1	0	0.00%	0.92%	0.00%	Carbamylation/Ala->Asn substitution	Addition of arginine due to transpeptidation/Addition of R	di-Methylation/Acetaldehyde +28/Ethylation/Ala->Val substitution/Cys->Met substitution
-1.9794	144	1046	54	1	5	0	0.69%	0.48%	0.00%	Thr->Val substitution		
-0.9474	257	1113	104	1	2	0	0.39%	0.18%	0.00%	Glu->Lys substitution		
228.1366	0	46	137	0	0	0	0.00%	0.00%	0.00%	dihydroxy/Pro->Glu substitution	Ala->Gly substitution/Glu->Asp substitution/Leu/Ile->Val substitution/Thr->Ser substitution/Gln->Asn substitution	Unannotated mass-shift 210.1616
58.0246	149	80	21	0	1	0	0.00%	1.25%	0.00%	First isotopic peak	Iodoacetamide derivative/Ala->Gln substitution/Gly->Asn substitution/Addition of Glycine/Addition of G	
-226.169	39	39	12	0	0	0	0.00%	0.00%	0.00%	Deletion of I/Deletion of L	Deletion of I/Deletion of L	
71.0372	22	111	29	1	2	0	4.55%	1.80%	0.00%	Acrylamide adduct/Gly->Gln substitution/Addition of A		
-488.2272	2	12	5	1	0	0	50.00%	0.00%	0.00%	Unannotated mass-shift -488.2272		
-147.0692	32	125	31	2	3	0	6.25%	2.40%	0.00%	Deletion of F		
185.1162	13	147	20	0	0	0	0.00%	0.00%	0.00%	Addition of lysine due to transpeptidation/Addition of K	Iodoacetamide derivative/Ala->Gln substitution/Gly->Asn substitution/Addition of Glycine/Addition of G	
12.0356	10	50	9	1	1	1	10.00%	2.00%	11.11%	Thr->Leu/Ile substitution/Ser->Val substitution		
-362.1626	0	14	5	0	1	0	0.00%	7.14%	0.00%	Unannotated mass-shift -229.1430	Removal of initiator methionine from protein N-terminus/Deletion of M	Thr->Val substitution
-0.047	287	1111	244	2	8	2	0.70%	0.72%	0.82%	Deamidation/Asn->Asp substitution/Gln->Glu substitution	Lysine oxidation to aminoadipic semialdehyde	
-200.0792	12	16	31	0	0	0	0.00%	0.00%	0.00%	Gln->Gly substitution/Deletion of A	Deletion of E	
38.9488	52	554	8	0	0	0	0.00%	0.00%	0.00%	First isotopic peak	Replacement of 2 protons by calcium	
-288.1254	5	24	24	0	0	0	0.00%	0.00%	0.00%	Amidation/Asp->Asn substitution/Glu->Gln substitution	Loss of arginine due to transpeptidation/Deletion of R	Removal of initiator methionine from protein N-terminus/Deletion of M
-225.1472	16	84	33	0	1	0	0.00%	1.19%	0.00%	dihydroxy/Pro->Glu substitution	Unannotated mass-shift -257.1376	
46.0414	71	95	3	0	1	0	0.00%	1.05%	0.00%	Acetylation/Ser->Glu substitution	tryptophan oxidation to kynurenin/Pro->Thr substitution	Gln->Lys substitution
-242.138	5	63	32	0	0	0	0.00%	0.00%	0.00%	Loss of C-terminal K from Heavy Chain of MAb/Deletion of K	Deletion of N	
-256.1538	9	65	34	0	0	0	0.00%	0.00%	0.00%	Deamidation/Asn->Asp substitution/Gln->Glu substitution	Unannotated mass-shift -257.1376	

213.123	12	89	20	0	2	0	0.00%	2.25%	0.00%	Iodoacetamide derivative/Ala->Gln substitution/Gly->Asn substitution/Addition of Glycine/Addition of G	Addition of arginine due to transpeptidation/Addition of R	
-127.0926	68	750	27	0	1	0	0.00%	0.13%	0.00%	First isotopic peak	Loss of C-terminal K from Heavy Chain of MAb/Deletion of K	
47.9848	30	73	70	0	0	0	0.00%	0.00%	0.00%	cysteine oxidation to cysteic acid		
-144.09	19	104	41	1	0	0	5.26%	0.00%	0.00%	Loss of C-terminal K from Heavy Chain of MAb/Deletion of K	reduction/Ser->Ala substitution/Tyr->Phe substitution	
-259.1358	5	38	17	0	1	0	0.00%	2.63%	0.00%	Loss of C-terminal K from Heavy Chain of MAb/Deletion of K	Removal of initiator methionine from protein N-terminus/Deletion of M	
30.0106	31	247	19	0	2	0	0.00%	0.81%	0.00%	hydroxymethyl/Ala->Thr substitution/Gly->Ser substitution		
0.959	276	1332	305	0	6	0	0.00%	0.45%	0.00%	Leu/Ile->Asn substitution		
1.9584	248	989	147	0	5	0	0.00%	0.51%	0.00%	Lysine oxidation to aminoadipic semialdehyde	glycosylated asparagine 18O labeling/Deamidation in presence of O18	
-369.238	8	31	21	0	0	0	0.00%	0.00%	0.00%	Deamidation/Asn->Asp substitution/Gln->Glu substitution	Deletion of I/Deletion of L	Unannotated mass-shift -257.1376
-228.1112	32	66	20	2	0	0	6.25%	0.00%	0.00%	Deletion of I/Deletion of L	Deletion of D	
17.9982	127	426	5	0	2	0	0.00%	0.47%	0.00%	Addition of lysine due to transpeptidation/Addition of K	Deletion of I/Deletion of L	glycosylated asparagine 18O labeling/Deamidation in presence of O18
26.0156	12	322	31	1	0	0	8.33%	0.00%	0.00%	Acetaldehyde +26/Ala->Pro substitution		
-14.9914	53	1082	5	1	3	0	1.89%	0.28%	0.00%	First isotopic peak	reduction/Ser->Ala substitution/Tyr->Phe substitution	
272.1274	1	34	97	0	0	0	0.00%	0.00%	0.00%	Acetylation/Ser->Glu substitution	Methylation/Asp->Glu substitution/Gly->Ala substitution/Ser->Thr substitution/Val->Leu/Ile substitution/Asn->Gln substitution	Unannotated mass-shift 216.1002
253.1534	6	79	20	0	0	0	0.00%	0.00%	0.00%	Addition of arginine due to transpeptidation/Addition of R	Acrylamide adduct/Gly->Gln substitution/Addition of A	Acetaldehyde +26/Ala->Pro substitution
22.9826	53	1030	12	0	1	0	0.00%	0.10%	0.00%	First isotopic peak	Unannotated mass-shift 23.9580	Thr->Val substitution
285.144	20	110	21	0	0	0	0.00%	0.00%	0.00%	Carbamylation/Ala->Asn substitution	Unannotated mass-shift 241.1788	Leu/Ile->Asn substitution
-215.0912	10	23	6	0	0	0	0.00%	0.00%	0.00%	Deletion of S	Deletion of Q	
248.1986	1	2	56	0	0	0	0.00%	0.00%	0.00%	Unannotated mass-shift 248.1986		
45.0108	124	1283	1	1	1	0	0.81%	0.08%	0.00%	First isotopic peak	S-Ethylcystine from Serine/Ser->Met substitution	
188.1052	1	24	93	0	0	0	0.00%	0.00%	0.00%	Carbamylation/Ala->Asn substitution	Pyro-glu from Q/Loss of ammonia	Unannotated mass-shift 162.1258
-186.1006	21	37	13	1	0	0	4.76%	0.00%	0.00%	Deletion of V	Deletion of S	
-214.1312	38	50	12	0	0	0	0.00%	0.00%	0.00%	Carbamylation/Ala->Asn substitution	Unannotated mass-shift -257.1376	
257.1372	15	214	12	0	1	0	0.00%	0.47%	0.00%	Acetylation/Ser->Glu substitution	Unannotated mass-shift 215.1270	
1.0674	240	1321	167	0	3	0	0.00%	0.23%	0.00%	Second isotopic peak	Unannotated mass-shift 162.1258	Deletion of Y
29.9736	9	41	6	0	2	0	0.00%	4.88%	0.00%	quinone/Val->Glu substitution		
159.933	29	28	46	0	0	0	0.00%	0.00%	0.00%	pyrophosphorylation of Ser/Thr		
199.1316	12	185	8	0	0	0	0.00%	0.00%	0.00%	Addition of lysine due to transpeptidation/Addition of K	Acrylamide adduct/Gly->Gln substitution/Addition of A	
-370.2218	5	46	21	0	0	0	0.00%	0.00%	0.00%	Deletion of I/Deletion of L	Unannotated mass-shift -257.1376	
-343.1856	1	28	13	0	0	0	0.00%	0.00%	0.00%	Unannotated mass-shift -229.1430	Deletion of N	
-270.1694	7	12	22	0	0	0	0.00%	0.00%	0.00%	Deamidation/Asn->Asp substitution/Gln->Glu substitution	Ala->Gly substitution/Glu->Asp substitution/Leu/Ile->Val substitution/Thr->Ser substitution/Gln->Asn substitution	Unannotated mass-shift -257.1376
230.0724	15	7	32	0	1	0	0.00%	14.29%	0.00%	Acetylation/Ser->Glu substitution	Iodoacetamide derivative/Ala->Gln substitution/Gly->Asn substitution/Addition of Glycine/Addition of G	Addition of M
81.9712	93	405	0	0	0	0	0.00%	0.00%	0.00%	Second isotopic peak	Phosphorylation	
-116.0584	3	247	12	0	0	0	0.00%	0.00%	0.00%	Loss of C-terminal K from Heavy Chain of MAb/Deletion of K	Thr->Leu/Ile substitution/Ser->Val substitution	

271.1282	10	87	18	0	0	0	0.00%	0.00%	0.00%	Addition of lysine due to transpeptidation/Addition of K	Unannotated mass-shift 306.0952	Deletion of Y
116.0626	0	394	2	0	0	0	0.00%	0.00%	0.00%	Replacement of 2 protons by calcium	Unannotated mass-shift 241.1788	Deletion of Y
0.883	122	578	184	0	3	0	0.00%	0.52%	0.00%	Unannotated mass-shift 0.8830		
279.9904	20	0	0	0	0	0	0.00%	0.00%	0.00%	Unannotated mass-shift 249.9810	hydroxymethyl/Ala->Thr substitution/Gly->Ser substitution	
189.0454	21	43	21	0	0	0	0.00%	0.00%	0.00%	Acetylation/Ser->Glu substitution	Oxidation or Hydroxylation/Ala->Ser substitution/Phe->Tyr substitution	Addition of M
129.0418	6	69	38	0	3	0	0.00%	4.35%	0.00%	monoglutamyl/Nmethylmaleimidehydrolysis/Addition of E		
-342.2264	4	32	21	0	0	0	0.00%	0.00%	0.00%	Deletion of I/Deletion of L	Unannotated mass-shift -229.1430	
162.0524	11	181	22	0	0	0	0.00%	0.00%	0.00%	Hexose		
-312.2156	6	41	30	0	0	0	0.00%	0.00%	0.00%	Oxidation or Hydroxylation/Ala->Ser substitution/Phe->Tyr substitution	Unannotated mass-shift -328.2110	
-1.0678	223	1015	149	0	1	1	0.00%	0.10%	0.67%	2-amino-3-oxo-butanoic_acid/Val->Pro substitution	Lys->Glu substitution	
-96.1058	30	15	22	0	0	0	0.00%	0.00%	0.00%	Loss of C-terminal K from Heavy Chain of MAb/Deletion of K	dihydroxy/Pro->Glu substitution	
-260.1532	13	22	11	0	1	0	0.00%	4.55%	0.00%	Deletion of I/Deletion of L	Deletion of F	
-0.959	280	1265	82	0	3	1	0.00%	0.24%	1.22%	Asn->Leu/Ile substitution		
-75.1822	0	27	78	0	0	0	0.00%	0.00%	0.00%	Unannotated mass-shift 52.9128	Loss of C-terminal K from Heavy Chain of MAb/Deletion of K	
-289.0734	0	11	8	0	0	0	0.00%	0.00%	0.00%	Unannotated mass-shift -289.0734		
156.0314	0	10	23	0	0	1	0.00%	0.00%	4.35%	Loss of C-terminal K from Heavy Chain of MAb/Deletion of K	Unannotated mass-shift 284.1268	
229.1424	9	182	4	0	2	0	0.00%	1.10%	0.00%	Methylation/Asp->Glu substitution/Gly->Ala substitution/Ser->Thr substitution/Val->Leu/Ile substitution/Asn->Gln substitution	Unannotated mass-shift 215.1270	
174.0884	1	19	74	0	0	0	0.00%	0.00%	0.00%	Thr->Leu/Ile substitution/Ser->Val substitution	Hexose	
252.9798	28	2	2	0	0	0	0.00%	0.00%	0.00%	Unannotated mass-shift 252.9798		
256.1904	9	159	6	0	0	0	0.00%	0.00%	0.00%	Addition of lysine due to transpeptidation/Addition of K	Addition of lysine due to transpeptidation/Addition of K	
200.1054	4	39	90	0	0	0	0.00%	0.00%	0.00%	reduction/Ser->Ala substitution/Tyr->Phe substitution	Unannotated mass-shift 216.1002	
262.141	1	17	102	0	0	0	0.00%	0.00%	0.00%	Unannotated mass-shift 234.0742	di-Methylation/Acetaldehyde +28/Ethylation/Ala->Val substitution/Cys->Met substitution	Gln->Lys substitution
168.0784	1	9	62	0	1	1	0.00%	11.11%	1.61%	Unannotated mass-shift 162.1258	Loss of arginine due to transpeptidation/Deletion of R	Hexose
71.0004	17	56	56	0	0	0	0.00%	0.00%	0.00%	Carbamylation/Ala->Asn substitution	Formylation/Ser->Asp substitution/Thr->Glu substitution	
152.9972	2	682	3	0	1	0	0.00%	0.15%	0.00%	First isotopic peak	Unannotated mass-shift 151.9956	
30.9832	6	106	149	0	0	0	0.00%	0.00%	0.00%	Formylation/Ser->Asp substitution/Thr->Glu substitution	glycosylated asparagine 18O labeling/Deamidation in presence of O18	
1.943	230	837	176	1	8	1	0.44%	0.96%	0.57%	Leu/Ile->Asp substitution		
130.1004	27	760	1	0	1	0	0.00%	0.13%	0.00%	Second isotopic peak	Addition of lysine due to transpeptidation/Addition of K	
242.1382	10	163	6	1	1	0	10.00%	0.61%	0.00%	Unannotated mass-shift 241.1788	Leu/Ile->Asn substitution	
-400.1888	1	10	2	0	0	0	0.00%	0.00%	0.00%	Unannotated mass-shift -400.1888		
42.0472	32	90	6	0	2	0	0.00%	2.22%	0.00%	tri-Methylation/Gly->Val substitution/Ala->Leu/Ile substitution/Propyl		
-259.1	3	19	12	1	0	0	33.33%	0.00%	0.00%	Deletion of Q	Removal of initiator methionine from protein N-terminus/Deletion of M	
216.1374	0	33	93	0	0	0	0.00%	0.00%	0.00%	Unannotated mass-shift 216.1002	Gln->Lys substitution	
-275.1638	5	45	23	0	0	0	0.00%	0.00%	0.00%	Loss of C-terminal K from Heavy Chain of MAb/Deletion of K	Deletion of F	
-356.2062	18	55	20	0	1	0	0.00%	1.82%	0.00%	Deletion of V	Unannotated mass-shift -257.1376	
-227.1386	5	71	17	0	0	0	0.00%	0.00%	0.00%	Gln->Gly substitution/Deletion of A	Loss of arginine due to transpeptidation/Deletion of R	
255.1696	10	93	10	0	0	0	0.00%	0.00%	0.00%	Iodoacetamide derivative/Ala->Gln substitution/Gly->Asn substitution/Addition of Glycine/Addition of G	Addition of arginine due to transpeptidation/Addition of R	tri-Methylation/Gly->Val substitution/Ala->Leu/Ile substitution/Propyl

185.041	79	415	3	0	0	0	0.00%	0.00%	0.00%	Second isotopic peak	Aminoethylbenzenesulfonylation	
99.0684	15	60	10	4	1	0	26.67%	1.67%	0.00%	N-isopropylcarboxamidomethyl/Addition of V		
-291.1582	2	40	12	0	0	0	0.00%	0.00%	0.00%	Loss of C-terminal K from Heavy Chain of MAb/Deletion of K	Deletion of Y	
257.1486	9	98	7	0	0	0	0.00%	0.00%	0.00%	Carbamylation/Ala->Asn substitution	Amidation/Asp->Asn substitution/Glu->Gln substitution	Unannotated mass-shift 215.1270
306.168	0	9	66	0	0	0	0.00%	0.00%	0.00%	Carbamylation/Ala->Asn substitution	Sodium adduct	Unannotated mass-shift 241.1788
199.1068	41	109	2	0	0	0	0.00%	0.00%	0.00%	Carbamylation/Ala->Asn substitution	Addition of arginine due to transpeptidation/Addition of R	
226.041	35	111	1	0	0	0	0.00%	0.00%	0.00%	Carbamylation/Ala->Asn substitution	Aminoethylbenzenesulfonylation	
-457.2536	5	30	8	0	0	1	0.00%	0.00%	12.50%	Unannotated mass-shift -328.2110	Deletion of E	
-340.2476	10	26	14	0	0	0	0.00%	0.00%	0.00%	Loss of C-terminal K from Heavy Chain of MAb/Deletion of K	Deletion of I/Deletion of L	Deletion of V
115.0272	13	100	7	0	1	0	0.00%	1.00%	0.00%	Addition of D		
-355.222	10	43	20	0	0	0	0.00%	0.00%	0.00%	Asn->Met substitution	Unannotated mass-shift -372.2198	
243.133	12	135	14	0	0	0	0.00%	0.00%	0.00%	Loss of C-terminal K from Heavy Chain of MAb/Deletion of K	Addition of arginine due to transpeptidation/Addition of R	Unannotated mass-shift 215.1270
-15.0214	68	762	9	0	0	0	0.00%	0.00%	0.00%	Pyro-glu from Q/Loss of ammonia	Second isotopic peak	
242.1266	5	45	9	1	0	0	20.00%	0.00%	0.00%	Unannotated mass-shift 241.1788	Lys->Glu substitution	
39.9518	17	312	27	0	0	0	0.00%	0.00%	0.00%	Second isotopic peak	Replacement of 2 protons by calcium	
147.0352	11	23	13	0	0	0	0.00%	0.00%	0.00%	Oxidation or Hydroxylation/Ala->Ser substitution/Phe->Tyr substitution	Addition of M	
377.0452	1	628	0	0	0	0	0.00%	0.00%	0.00%	Addition of lysine due to transpeptidation/Addition of K	Unannotated mass-shift 249.9810	Lysine oxidation to aminoadipic semialdehyde
-298.1998	6	35	22	0	0	0	0.00%	0.00%	0.00%	Unannotated mass-shift -328.2110	hydroxymethyl/Ala->Thr substitution/Gly->Ser substitution	
312.202	21	67	9	0	0	0	0.00%	0.00%	0.00%	Addition of arginine due to transpeptidation/Addition of R	Addition of arginine due to transpeptidation/Addition of R	
163.0634	5	56	19	0	2	1	0.00%	3.57%	5.26%	Addition of Y		
311.1298	48	29	1	0	0	0	0.00%	0.00%	0.00%	Addition of lysine due to transpeptidation/Addition of K	Aminoethylbenzenesulfonylation	
300.1208	1	95	93	0	0	0	0.00%	0.00%	0.00%	Oxidation or Hydroxylation/Ala->Ser substitution/Phe->Tyr substitution	Unannotated mass-shift 284.1268	
-2.0526	153	717	44	1	2	0	0.65%	0.28%	0.00%	2-amino-3-oxo-butanoic_acid/Val->Pro substitution	Lys->Gln substitution	
18.032	107	475	12	0	1	0	0.00%	0.21%	0.00%	Unannotated mass-shift 3.0076	glycosylated asparagine 18O labeling/Deamidation in presence of O18	Thr->Leu/Ile substitution/Ser->Val substitution
-1.077	206	976	125	0	2	0	0.00%	0.21%	0.00%	Carbamylation/Ala->Asn substitution	Unannotated mass-shift 284.1268	Unannotated mass-shift -328.2110
282.1686	20	87	55	0	0	0	0.00%	0.00%	0.00%	Carbamylation/Ala->Asn substitution	Unannotated mass-shift 241.1788	2-amino-3-oxo-butanoic_acid/Val->Pro substitution
-0.9942	280	1795	255	6	22	4	2.14%	1.23%	1.57%	Deamidation/Asn->Asp substitution/Gln->Glu substitution	Thr->Val substitution	
-242.1268	12	45	20	0	0	0	0.00%	0.00%	0.00%	Deletion of I/Deletion of L	Deletion of E	
303.9924	94	58	0	0	0	0	0.00%	0.00%	0.00%	Second isotopic peak	Unannotated mass-shift 301.9864	
-218.0722	7	10	9	0	0	0	0.00%	0.00%	0.00%	Deletion of S	Removal of initiator methionine from protein N-terminus/Deletion of M	
256.1542	4	128	8	0	0	0	0.00%	0.00%	0.00%	amidination of lysines or N-terminal amines with methyl acetimidate/Ser->Gln substitution	Unannotated mass-shift 215.1270	
-440.239	2	25	4	0	0	0	0.00%	0.00%	0.00%	Unannotated mass-shift -440.2390		
-493.265	1	16	5	0	0	0	0.00%	0.00%	0.00%	tryptophan oxidation to kynurenin/Pro->Thr substitution	Asn->Gly substitution/Gln->Ala substitution/Deletion of G	Unannotated mass-shift -440.2390
-427.2434	7	37	12	0	0	0	0.00%	0.00%	0.00%	Oxidation or Hydroxylation/Ala->Ser substitution/Phe->Tyr substitution	Unannotated mass-shift -328.2110	Deletion of D
-227.1268	23	43	16	1	1	1	4.35%	2.33%	6.25%	Deletion of I/Deletion of L	Deletion of N	
259.135	4	123	8	0	0	0	0.00%	0.00%	0.00%	Addition of lysine due to transpeptidation/Addition of K	Addition of M	
-253.1536	7	62	14	0	0	0	0.00%	0.00%	0.00%	Oxidation or Hydroxylation/Ala->Ser substitution/Phe->Tyr substitution	Addition of M	Unannotated mass-shift -400.1888
-15.0112	23	480	28	1	8	3	4.35%	1.67%	10.71%	lactic acid from N-term Ser/Lys->Leu/Ile substitution/ISD (z+2)-series		
-137.0588	11	39	6	0	0	0	0.00%	0.00%	0.00%	Deletion of H		

-28.0314	13	93	14	0	0	0	0.00%	0.00%	0.00%	Val->Ala substitution/Met->Cys substitution		
-156.09	21	53	7	1	2	0	4.76%	3.77%	0.00%	Deletion of V	Asn->Gly substitution/Gln->Ala substitution/Deletion of G	
54.94	32	73	2	0	0	0	0.00%	0.00%	0.00%	Loss of C-terminal K from Heavy Chain of MAb/Deletion of K	Aminoethylbenzenesulfonylation	
462.098	0	562	0	0	0	0	0.00%	0.00%	0.00%	Unannotated mass-shift 420.0506	tri-Methylation/Gly->Val substitution/Ala->Leu/Ile substitution/Propyl	
-1.0894	181	855	138	1	0	0	0.55%	0.00%	0.00%	Unannotated mass-shift -1.0894		
-234.1008	6	26	6	0	0	0	0.00%	0.00%	0.00%	Gln->Gly substitution/Deletion of A	Deletion of Y	
1.9216	183	757	131	0	4	1	0.00%	0.53%	0.76%	First isotopic peak	Second isotopic peak	Unannotated mass-shift -1.0894
276.1214	41	8	3	0	0	0	0.00%	0.00%	0.00%	Unannotated mass-shift 234.0742	tri-Methylation/Gly->Val substitution/Ala->Leu/Ile substitution/Propyl	
256.1314	1	25	57	0	0	0	0.00%	0.00%	0.00%	Acetylation/Ser->Glu substitution	Unannotated mass-shift 216.1002	Thr->Val substitution
-2.0044	111	830	25	2	5	1	1.80%	0.60%	4.00%	Addition of arginine due to transpeptidation/Addition of R	Unannotated mass-shift -229.1430	Acrylamide adduct/Gly->Gln substitution/Addition of A
-243.1336	6	77	10	0	0	0	0.00%	0.00%	0.00%	Loss of arginine due to transpeptidation/Deletion of R	Deletion of S	
-258.0852	0	8	19	0	0	1	0.00%	0.00%	5.26%	Unannotated mass-shift -257.1376	Glu->Lys substitution	
26.0518	7	40	8	0	0	0	0.00%	0.00%	0.00%	Ser->Leu/Ile substitution		
-1.02	394	1782	164	1	19	0	0.25%	1.07%	0.00%	Amidation/Asp->Asn substitution/Glu->Gln substitution	Lys->Gln substitution	
370.22	14	115	53	0	0	0	0.00%	0.00%	0.00%	Unannotated mass-shift 241.1788	monoglutamyl/Nmethylmaleimidehydrolysis/Addition of E	
97.9664	23	27	31	0	0	0	0.00%	0.00%	0.00%	First isotopic peak	Phosphorylation	Asn->Met substitution
176.7462	0	9	27	0	0	0	0.00%	0.00%	0.00%	Unannotated mass-shift 176.7462		
-414.1388	2	2	1	0	0	0	0.00%	0.00%	0.00%	Unannotated mass-shift -414.1388		
-25.032	0	207	8	0	1	0	0.00%	0.48%	0.00%	dihydroxy/Pro->Glu substitution	Asn->Gly substitution/Gln->Ala substitution/Deletion of G	
-214.0956	26	33	10	0	0	0	0.00%	0.00%	0.00%	Deletion of V	Deletion of D	
176.1408	0	5	58	0	0	0	0.00%	0.00%	0.00%	Unannotated mass-shift 162.1258	Methylation/Asp->Glu substitution/Gly->Ala substitution/Ser->Thr substitution/Val->Leu/Ile substitution/Asn->Gln substitution	
72.0332	23	348	1	0	0	0	0.00%	0.00%	0.00%	Carbamylation/Ala->Asn substitution	Addition of lysine due to transpeptidation/Addition of K	Deletion of V
123.9124	0	1	2	0	0	0	0.00%	0.00%	0.00%	Carbamylation/Ala->Asn substitution	Unannotated mass-shift 52.9128	Formylation/Ser->Asp substitution/Thr->Glu substitution
-241.1428	12	47	20	0	0	0	0.00%	0.00%	0.00%	Oxidation or Hydroxylation/Ala->Ser substitution/Phe->Tyr substitution	Unannotated mass-shift -257.1376	
134.047	0	11	8	0	0	1	0.00%	0.00%	12.50%	3-methyl-2-pyridyl isocyanate		
15.0188	28	179	4	0	0	0	0.00%	0.00%	0.00%	First isotopic peak	Methylation/Asp->Glu substitution/Gly->Ala substitution/Ser->Thr substitution/Val->Leu/Ile substitution/Asn->Gln substitution	
24.9414	0	9	9	0	0	0	0.00%	0.00%	0.00%	Deamidation/Asn->Asp substitution/Gln->Glu substitution	Unannotated mass-shift 23.9580	
243.1226	8	124	13	0	0	0	0.00%	0.00%	0.00%	Addition of lysine due to transpeptidation/Addition of K	Addition of D	
406.1596	21	39	10	0	0	0	0.00%	0.00%	0.00%	HexNAc2		
66.9272	0	20	15	0	0	0	0.00%	0.00%	0.00%	Unannotated mass-shift 52.9128	Methylation/Asp->Glu substitution/Gly->Ala substitution/Ser->Thr substitution/Val->Leu/Ile substitution/Asn->Gln substitution	
204.083	43	121	2	0	0	1	0.00%	0.00%	50.00%	First isotopic peak	N-Acetylhexosamine	
-202.0598	7	25	7	0	0	0	0.00%	0.00%	0.00%	Deletion of S	Deletion of D	
-399.2122	4	23	7	1	0	0	25.00%	0.00%	0.00%	amidination of lysines or N-terminal amines with methyl acetimidate/Ser->Gln substitution	Unannotated mass-shift -440.2390	
-10.032	6	33	31	0	0	0	0.00%	0.00%	0.00%	First isotopic peak	Asn->Met substitution	Val->Ala substitution/Met->Cys substitution

1.0424	245	3032	188	0	5	0	0.00%	0.17%	0.00%	Unannotated mass-shift 162.1258	Ala->Gly substitution/Glu->Asp substitution/Leu/Ile->Val substitution/Thr->Ser substitution/Gln->Asn substitution	Deletion of F
345.0188	0	480	0	0	0	0	0.00%	0.00%	0.00%	Deamidation/Asn->Asp substitution/Gln->Glu substitution	Unannotated mass-shift 301.9864	tri-Methylation/Gly->Val substitution/Ala->Leu/Ile substitution/Propyl
284.9596	48	3	1	0	0	0	0.00%	0.00%	0.00%	Pyro-glu from Q/Loss of ammonia	Unannotated mass-shift 301.9864	
-172.0846	13	20	11	0	0	0	0.00%	0.00%	0.00%	Gln->Gly substitution/Deletion of A	Deletion of T	
-471.2496	6	12	5	0	0	0	0.00%	0.00%	0.00%	Deamidation/Asn->Asp substitution/Gln->Glu substitution	Oxidation or Hydroxylation/Ala->Ser substitution/Phe->Tyr substitution	Unannotated mass-shift -488.2272
-469.29	3	15	9	0	1	0	0.00%	6.67%	0.00%	Deletion of I/Deletion of L	Deletion of V	Unannotated mass-shift -257.1376
114.0436	17	42	10	0	0	1	0.00%	0.00%	10.00%	ubiquitinylation residue/Double Carbamidomethylation/Addition of N		
275.1632	8	88	5	0	0	0	0.00%	0.00%	0.00%	Pyro-glu from Q/Loss of ammonia	Carboxylation/Ala->Asp substitution	Unannotated mass-shift 248.1986
384.0816	3	465	2	0	0	0	0.00%	0.00%	0.00%	Addition of lysine due to transpeptidation/Addition of K	Unannotated mass-shift 3.0076	Unannotated mass-shift 252.9798
186.089	1	15	57	0	0	0	0.00%	0.00%	0.00%	Pyro-glu from Q/Loss of ammonia	N-Acetylhexosamine	Gln->Lys substitution
17.9552	7	62	15	0	2	0	0.00%	3.23%	0.00%	Leu/Ile->Met substitution		
-198.1372	19	20	2	0	0	0	0.00%	0.00%	0.00%	Deletion of V	Deletion of V	
236.0678	17	50	14	0	0	0	0.00%	0.00%	0.00%	Phosphorylation	Addition of arginine due to transpeptidation/Addition of R	
-0.9414	223	914	76	0	2	0	0.00%	0.22%	0.00%	First isotopic peak	Amidation/Asp->Asn substitution/Glu->Gln substitution	Asn->Leu/Ile substitution
105.8246	0	5	32	0	0	0	0.00%	0.00%	0.00%	Unannotated mass-shift 52.9128	Unannotated mass-shift 52.9128	
284.1838	17	83	2	0	0	0	0.00%	0.00%	0.00%	Carbamylation/Ala->Asn substitution	Unannotated mass-shift 241.1788	
-325.163	1	9	4	0	0	0	0.00%	0.00%	0.00%	Unannotated mass-shift -488.2272	Addition of Y	
359.0082	4	1	33	0	0	0	0.00%	0.00%	0.00%	Unannotated mass-shift 52.9128	Unannotated mass-shift 306.0952	
1.8778	104	374	85	0	1	1	0.00%	0.27%	1.18%	Pyro-glu from Q/Loss of ammonia	Lys->Glu substitution	Leu/Ile->Met substitution
-313.164	4	29	16	0	1	0	0.00%	3.45%	0.00%	Oxidation or Hydroxylation/Ala->Ser substitution/Phe->Tyr substitution	Unannotated mass-shift -328.2110	Glu->Lys substitution
-1.0416	303	1214	128	1	2	0	0.33%	0.17%	0.00%	Pyro-glu from Q/Loss of ammonia	Second isotopic peak	proline oxidation to pyroglutamic acid/Tryptophan oxidation to oxolactone/Thr->Asp substitution
-185.116	7	64	25	0	0	0	0.00%	0.00%	0.00%	Loss of C-terminal K from Heavy Chain of MAb/Deletion of K	Asn->Gly substitution/Gln->Ala substitution/Deletion of G	
-411.2848	6	21	6	0	0	0	0.00%	0.00%	0.00%	Oxidation or Hydroxylation/Ala->Ser substitution/Phe->Tyr substitution	Deletion of V	Unannotated mass-shift -328.2110
14.9734	37	69	7	0	1	0	0.00%	1.45%	0.00%	Leu/Ile->Gln substitution/Val->Asn substitution		
0.8474	68	312	98	1	1	1	1.47%	0.32%	1.02%	Lys->Gln substitution	Unannotated mass-shift 0.8830	
326.1936	13	83	52	0	0	0	0.00%	0.00%	0.00%	Acetylation/Ser->Glu substitution	Carbamylation/Ala->Asn substitution	Unannotated mass-shift 241.1788
-30.0104	9	82	15	0	0	1	0.00%	0.00%	6.67%	Proline oxidation to pyrrolidinone/Ser->Gly substitution/Thr->Ala substitution		
-356.2174	9	44	12	0	0	0	0.00%	0.00%	0.00%	Acetylation/Ser->Glu substitution	Acetylation/Ser->Glu substitution	Unannotated mass-shift -440.2390
-410.2892	3	8	2	0	0	0	0.00%	0.00%	0.00%	Unannotated mass-shift -440.2390	Leu/Ile->Gln substitution/Val->Asn substitution	Leu/Ile->Gln substitution/Val->Asn substitution
-13.99	13	790	2	0	1	0	0.00%	0.13%	0.00%	Second isotopic peak	reduction/Ser->Ala substitution/Tyr->Phe substitution	
272.1632	0	26	55	0	0	0	0.00%	0.00%	0.00%	dihydroxy/Pro->Glu substitution	Unannotated mass-shift 210.1616	hydroxymethyl/Ala->Thr substitution/Gly->Ser substitution
-256.1166	2	14	5	0	0	0	0.00%	0.00%	0.00%	Deletion of Q	Deletion of Q	
-358.1856	5	21	6	0	1	1	0.00%	4.76%	16.67%	Unannotated mass-shift -229.1430	Deletion of E	
0.0728	190	1102	120	0	1	0	0.00%	0.09%	0.00%	Gln->Lys substitution	Gln->Lys substitution	
355.904	43	0	0	0	0	0	0.00%	0.00%	0.00%	Unannotated mass-shift 301.9864	Replacement of 2 protons by iron	
-160.0308	10	58	10	1	0	0	10.00%	0.00%	0.00%	Deletion of C		
-184.085	5	30	13	0	0	0	0.00%	0.00%	0.00%	Oxidation or Hydroxylation/Ala->Ser substitution/Phe->Tyr substitution	Gln->Gly substitution/Deletion of A	Deletion of E

-1.9676	128	856	37	0	4	0	0.00%	0.47%	0.00%	Amidation/Asp->Asn substitution/Glu->Gln substitution	Amidation/Asp->Asn substitution/Glu->Gln substitution	
-399.2488	4	28	13	0	1	0	0.00%	3.57%	0.00%	Gln->Gly substitution/Deletion of A	Unannotated mass-shift -328.2110	
-17.0088	65	550	32	1	2	0	1.54%	0.36%	0.00%	First isotopic peak	Dehydration/Pyro-glu from E	
-441.294	3	15	15	0	0	0	0.00%	0.00%	0.00%	Deletion of I/Deletion of L	Unannotated mass-shift -328.2110	
-244.0694	10	35	16	0	0	0	0.00%	0.00%	0.00%	Deletion of E	Deletion of D	
-213.1482	1	18	4	0	0	0	0.00%	0.00%	0.00%	Oxidation or Hydroxylation/Ala->Ser substitution/Phe->Tyr substitution	Unannotated mass-shift -229.1430	
36.919	0	21	63	0	0	0	0.00%	0.00%	0.00%	Unannotated mass-shift 52.9128	reduction/Ser->Ala substitution/Tyr->Phe substitution	
-318.1366	2	4	1	0	0	0	0.00%	0.00%	0.00%	Acetylation/Ser->Glu substitution	Lys->Gly substitution	Unannotated mass-shift -289.0734
-328.1748	3	27	22	0	1	0	0.00%	3.70%	0.00%	Gln->Gly substitution/Deletion of A	Unannotated mass-shift -257.1376	
275.13	8	15	2	0	0	0	0.00%	0.00%	0.00%	Carbamylation/Ala->Asn substitution	Asn->Met substitution	Unannotated mass-shift 215.1270
414.2462	7	82	57	0	1	0	0.00%	1.22%	0.00%	First isotopic peak	N-Acetylhexosamine	Unannotated mass-shift 210.1616
225.1482	3	31	3	0	0	0	0.00%	0.00%	0.00%	Addition of lysine due to transpeptidation/Addition of K	2-amino-3-oxo-butanoic_acid/Val->Pro substitution	N-isopropylcarboxamidomethyl/Addition of V
-413.2278	7	38	10	0	0	0	0.00%	0.00%	0.00%	Ser->Asn substitution/Thr->Gln substitution	Unannotated mass-shift -440.2390	
-414.2116	4	23	8	1	1	0	25.00%	4.35%	0.00%	2-amino-3-oxo-butanoic_acid/Val->Pro substitution	Leu/Ile->Asp substitution	Unannotated mass-shift -414.1388
-264.1116	2	13	2	0	0	0	0.00%	0.00%	0.00%	Deletion of Y	Deletion of T	
-273.1136	5	29	10	0	0	0	0.00%	0.00%	0.00%	Deletion of I/Deletion of L	Deletion of C	
239.1268	14	131	1	0	0	0	0.00%	0.00%	0.00%	Acetylation/Ser->Glu substitution	Dehydration/Pyro-glu from E	Unannotated mass-shift 215.1270
291.1582	3	73	2	0	1	0	0.00%	1.37%	0.00%	Addition of lysine due to transpeptidation/Addition of K	Addition of Y	
-414.2594	1	5	2	0	0	0	0.00%	0.00%	0.00%	Acetylation/Ser->Glu substitution	Unannotated mass-shift -328.2110	Deletion of Q
58.0054	10	70	16	2	0	0	20.00%	0.00%	0.00%	Iodoacetic acid derivative/Ala->Glu substitution/Gly->Asp substitution		
-32.0076	11	24	31	0	0	0	0.00%	0.00%	0.00%	Acetylation/Ser->Glu substitution	Iodoacetamide derivative/Ala->Gln substitution/Gly->Asn substitution/Addition of Glycine/Addition of G	Removal of initiator methionine from protein N-terminus/Deletion of M
-173.1056	2	9	2	0	0	0	0.00%	0.00%	0.00%	Carbamylation/Ala->Asn substitution	amidination of lysines or N-terminal amines with methyl acetimidate/Ser->Gln substitution	Unannotated mass-shift -257.1376
-89.0296	11	41	11	0	0	0	0.00%	0.00%	0.00%	Removal of initiator methionine from protein N-terminus, then acetylation of the new N-terminus		
-97.0534	6	15	6	1	0	0	16.67%	0.00%	0.00%	Deletion of P		
366.0698	4	160	69	0	0	1	0.00%	0.00%	1.45%	Aminoethylbenzenesulfonylation	Aminoethylbenzenesulfonylation	
-1.9318	119	571	19	0	1	0	0.00%	0.18%	0.00%	Amidation/Asp->Asn substitution/Glu->Gln substitution	Glu->Lys substitution	
-314.159	7	40	13	1	1	1	14.29%	2.50%	7.69%	Unannotated mass-shift -257.1376	Asn->Gly substitution/Gln->Ala substitution/Deletion of G	
-386.18	2	24	13	0	1	0	0.00%	4.17%	0.00%	Unannotated mass-shift -257.1376	Deletion of E	
496.3114	0	14	1	0	1	0	0.00%	7.14%	0.00%	Unannotated mass-shift 496.3114		
284.1606	3	59	13	0	0	0	0.00%	0.00%	0.00%	Acetylation/Ser->Glu substitution	dihydroxy/Pro->Glu substitution	Unannotated mass-shift 210.1616
-29.9922	6	43	6	0	0	0	0.00%	0.00%	0.00%	Homoserine/Met->Thr substitution		
-112.0792	15	120	4	0	0	0	0.00%	0.00%	0.00%	First isotopic peak	Deletion of I/Deletion of L	
-12.0356	3	56	6	0	0	0	0.00%	0.00%	0.00%	Leu/Ile->Thr substitution/Val->Ser substitution		
-435.1622	0	9	3	0	0	0	0.00%	0.00%	0.00%	Unannotated mass-shift -435.1622		
-1.1218	98	500	90	0	0	0	0.00%	0.00%	0.00%	Acetylation/Ser->Glu substitution	Replacement of 2 protons by iron	Deletion of P
-185.08	15	19	6	0	0	0	0.00%	0.00%	0.00%	Gln->Gly substitution/Deletion of A	Deletion of N	
212.1414	0	19	64	0	0	1	0.00%	0.00%	1.56%	Oxidation or Hydroxylation/Ala->Ser substitution/Phe->Tyr substitution	Ala->Gly substitution/Glu->Asp substitution/Leu/Ile->Val substitution/Thr->Ser substitution/Gln->Asn substitution	Unannotated mass-shift 210.1616

-213.1114	18	21	10	0	1	0	0.00%	4.76%	0.00%	Deletion of V	Deletion of N	
0.0842	172	837	98	0	0	0	0.00%	0.00%	0.00%	replacement of proton with ammonium ion	reduction/Ser->Ala substitution/Tyr->Phe substitution	Glu->Lys substitution
-186.0794	22	31	8	0	0	0	0.00%	0.00%	0.00%	Deletion of W		
230.1152	0	14	75	0	0	0	0.00%	0.00%	0.00%	Methylation/Asp->Glu substitution/Gly->Ala substitution/Ser->Thr substitution/Val->Leu/Ile substitution/Asn->Gln substitution	Unannotated mass-shift 216.1002	
214.1206	3	30	83	0	0	0	0.00%	0.00%	0.00%	Unannotated mass-shift 216.1002	Thr->Val substitution	
184.1326	5	19	1	0	0	0	0.00%	0.00%	0.00%	Addition of arginine due to transpeptidation/Addition of R	di-Methylation/Acetaldehyde +28/Ethylation/Ala->Val substitution/Cys->Met substitution	
-471.2796	4	16	4	1	0	0	25.00%	0.00%	0.00%	Pyro-glu from Q/Loss of ammonia	Ala->Gly substitution/Glu->Asp substitution/Leu/Ile->Val substitution/Thr->Ser substitution/Gln->Asn substitution	Unannotated mass-shift -440.2390
208.0616	11	33	9	0	0	0	0.00%	0.00%	0.00%	Addition of lysine due to transpeptidation/Addition of K	Phosphorylation	
190.157	0	42	44	0	0	0	0.00%	0.00%	0.00%	Unannotated mass-shift 162.1258	di-Methylation/Acetaldehyde +28/Ethylation/Ala->Val substitution/Cys->Met substitution	
2.9528	78	325	27	0	3	1	0.00%	0.92%	3.70%	Second isotopic peak	Lys->Glu substitution	
99.0318	6	9	15	0	0	0	0.00%	0.00%	0.00%	EGS crosslinker to Lys or N-terminus following hydroxylamine cleavage		
303.17	5	55	11	0	0	0	0.00%	0.00%	0.00%	Acetylation/Ser->Glu substitution	Unannotated mass-shift 162.1258	EGS crosslinker to Lys or N-terminus following hydroxylamine cleavage
285.2058	20	276	8	0	0	0	0.00%	0.00%	0.00%	Oxidation or Hydroxylation/Ala->Ser substitution/Phe->Tyr substitution	Unannotated mass-shift 241.1788	di-Methylation/Acetaldehyde +28/Ethylation/Ala->Val substitution/Cys->Met substitution
158.108	15	359	0	0	0	0	0.00%	0.00%	0.00%	Second isotopic peak	Addition of arginine due to transpeptidation/Addition of R	
160.9364	18	24	4	0	0	0	0.00%	0.00%	0.00%	First isotopic peak	pyrophosphorylation of Ser/Thr	
-354.2628	7	28	14	0	0	0	0.00%	0.00%	0.00%	Unannotated mass-shift -372.2198	Leu/Ile->Met substitution	
-344.2062	1	9	7	0	0	0	0.00%	0.00%	0.00%	reduction/Ser->Ala substitution/Tyr->Phe substitution	Unannotated mass-shift -328.2110	
-16.0416	29	220	4	0	1	0	0.00%	0.46%	0.00%	Deamidation/Asn->Asp substitution/Gln->Glu substitution	Pyro-glu from Q/Loss of ammonia	
-435.1958	0	8	5	0	0	0	0.00%	0.00%	0.00%	Loss of C-terminal K from Heavy Chain of MAb/Deletion of K	Deletion of F	Deletion of C
458.273	4	70	59	0	0	0	0.00%	0.00%	0.00%	Unannotated mass-shift 241.1788	Methylation/Asp->Glu substitution/Gly->Ala substitution/Ser->Thr substitution/Val->Leu/Ile substitution/Asn->Gln substitution	N-Acetylhexosamine
-282.169	2	17	14	0	0	0	0.00%	0.00%	0.00%	Loss of C-terminal K from Heavy Chain of MAb/Deletion of K	Asn->Gly substitution/Gln->Ala substitution/Deletion of G	Deletion of P
19.0316	32	352	4	0	0	0	0.00%	0.00%	0.00%	Second isotopic peak	replacement of proton with ammonium ion	
-285.1692	4	17	8	0	0	0	0.00%	0.00%	0.00%	Unannotated mass-shift -257.1376	Val->Ala substitution/Met->Cys substitution	
-369.2002	5	21	9	0	0	0	0.00%	0.00%	0.00%	Acrylamide adduct/Gly->Gln substitution/Addition of A	Unannotated mass-shift -440.2390	
61.9116	1	9	47	0	0	0	0.00%	0.00%	0.00%	Replacement of 2 protons by zinc		
355.906	39	0	0	0	0	0	0.00%	0.00%	0.00%	Unannotated mass-shift 301.9864	Replacement of 2 protons by iron	
-99.079	5	29	2	0	0	0	0.00%	0.00%	0.00%	Arg->Gly substitution		
-199.0958	11	33	14	0	0	0	0.00%	0.00%	0.00%	Gln->Gly substitution/Deletion of A	Deletion of Q	
-402.1572	0	7	9	0	0	0	0.00%	0.00%	0.00%	Deletion of I/Deletion of L	Unannotated mass-shift -289.0734	
-300.144	1	16	8	0	0	0	0.00%	0.00%	0.00%	Loss of C-terminal K from Heavy Chain of MAb/Deletion of K	Asn->Gly substitution/Gln->Ala substitution/Deletion of G	Deletion of D
59.0002	23	86	3	0	1	0	0.00%	1.16%	0.00%	Carbamylation/Ala->Asn substitution	Oxidation or Hydroxylation/Ala->Ser substitution/Phe->Tyr substitution	
-271.153	4	16	9	0	0	0	0.00%	0.00%	0.00%	Ala->Gly substitution/Glu->Asp substitution/Leu/Ile->Val substitution/Thr->Ser substitution/Gln->Asn substitution	Unannotated mass-shift -257.1376	

254.1158	1	4	58	0	0	0	0.00%	0.00%	0.00%	Unannotated mass-shift 284.1268	Proline oxidation to pyrrolidinone/Ser->Gly substitution/Thr->Ala substitution	
371.1808	1	26	0	0	2	0	0.00%	7.69%	0.00%	Acetylation/Ser->Glu substitution	Unannotated mass-shift 215.1270	ubiquitinylation residue/Double Carbamidomethylation/Addition of N
212.1628	1	27	0	0	1	0	0.00%	3.70%	0.00%	Iodoacetamide derivative/Ala->Gln substitution/Gly->Asn substitution/Addition of Glycine/Addition of G	Addition of arginine due to transpeptidation/Addition of R	Asn->Leu/Ile substitution
-0.0688	257	1078	233	1	5	0	0.39%	0.46%	0.00%	Deamidation/Asn->Asp substitution/Gln->Glu substitution	Gln->Lys substitution	Unannotated mass-shift -1.0894
272.1484	5	56	4	0	0	0	0.00%	0.00%	0.00%	Iodoacetamide derivative/Ala->Gln substitution/Gly->Asn substitution/Addition of Glycine/Addition of G	Unannotated mass-shift 215.1270	
0.0902	166	774	78	0	0	0	0.00%	0.00%	0.00%	First isotopic peak	Gln->Lys substitution	Glu->Lys substitution
-285.1316	11	21	10	0	0	0	0.00%	0.00%	0.00%	Addition of lysine due to transpeptidation/Addition of K	Ser->Asn substitution/Thr->Gln substitution	Unannotated mass-shift -440.2390
-298.1636	5	15	8	0	1	0	0.00%	6.67%	0.00%	Carbamylation/Ala->Asn substitution	N-isopropylcarboxamidomethyl/Addition of V	Unannotated mass-shift -440.2390
-330.1728	2	14	10	0	0	0	0.00%	0.00%	0.00%	Unannotated mass-shift -372.2198	tri-Methylation/Gly->Val substitution/Ala->Leu/Ile substitution/Propyl	
421.0354	0	405	0	0	0	0	0.00%	0.00%	0.00%	Deamidation/Asn->Asp substitution/Gln->Glu substitution	Unannotated mass-shift 420.0506	
-0.0576	270	1124	294	1	5	0	0.37%	0.45%	0.00%	Pyro-glu from Q/Loss of ammonia	proline oxidation to pyroglutamic acid/Tryptophan oxidation to oxolactone/Thr->Asp substitution	glycosylated asparagine 18O labeling/Deamidation in presence of O18
103.9246	0	8	29	0	0	0	0.00%	0.00%	0.00%	Acetylation/Ser->Glu substitution	Replacement of 2 protons by zinc	
0.1282	98	385	33	0	1	0	0.00%	0.26%	0.00%	Unannotated mass-shift 0.1282		
-216.0752	5	35	21	0	0	0	0.00%	0.00%	0.00%	Deletion of E	Deletion of S	
-313.2008	2	28	10	0	0	0	0.00%	0.00%	0.00%	Carbamylation/Ala->Asn substitution	Deletion of V	Unannotated mass-shift -257.1376
166.0088	22	24	0	0	0	0	0.00%	0.00%	0.00%	Pyro-glu from Q/Loss of ammonia	Aminoethylbenzenesulfonylation	
121.977	4	31	6	0	0	0	0.00%	0.00%	0.00%	Acetylation/Ser->Glu substitution	Phosphorylation	
-442.2418	0	13	8	0	1	0	0.00%	7.69%	0.00%	Deletion of I/Deletion of L	Unannotated mass-shift -328.2110	Glu->Lys substitution
-300.18	4	19	12	0	0	0	0.00%	0.00%	0.00%	Gln->Gly substitution/Deletion of A	Unannotated mass-shift -229.1430	
70.0426	13	106	5	0	0	0	0.00%	0.00%	0.00%	Crotonaldehyde/Butyryl		
-486.244	7	16	8	0	0	0	0.00%	0.00%	0.00%	First isotopic peak	Deletion of D	Unannotated mass-shift -372.2198
-296.279	0	1	1	0	0	0	0.00%	0.00%	0.00%	Unannotated mass-shift -296.2790		
-475.278	3	10	4	0	0	0	0.00%	0.00%	0.00%	Unannotated mass-shift -328.2110	Deletion of F	
-1.8774	56	256	6	0	0	0	0.00%	0.00%	0.00%	Unannotated mass-shift -1.8774		
29.0148	12	136	1	0	1	0	0.00%	0.74%	0.00%	deamidation followed by esterification with ethanol		
242.1822	8	195	1	0	1	0	0.00%	0.51%	0.00%	First isotopic peak	Unannotated mass-shift 241.1788	
181.04	26	7	0	0	0	0	0.00%	0.00%	0.00%	Second isotopic peak	Phosphorylation	N-isopropylcarboxamidomethyl/Addition of V
71.985	26	14	3	0	1	0	0.00%	7.14%	0.00%	Acetylation/Ser->Glu substitution	quinone/Val->Glu substitution	
-357.1542	2	19	9	1	0	1	50.00%	0.00%	11.11%	Acetylation/Ser->Glu substitution	Unannotated mass-shift -414.1388	Leu/Ile->Gln substitution/Val->Asn substitution
249.9586	1	5	1	0	0	0	0.00%	0.00%	0.00%	Unannotated mass-shift 52.9128	glycerylphosphorylethanolamine	
-443.202	2	30	14	0	0	1	0.00%	0.00%	7.14%	Acetylation/Ser->Glu substitution	Lys->Gly substitution	Unannotated mass-shift -414.1388
3.0342	73	1183	11	0	2	0	0.00%	0.17%	0.00%	tryptophan oxidation to kynurenin/Pro->Thr substitution	Asn->Leu/Ile substitution	
319.0116	35	1	0	0	0	0	0.00%	0.00%	0.00%	Unannotated mass-shift 301.9864	replacement of proton with ammonium ion	
-0.1738	43	162	55	0	2	0	0.00%	1.24%	0.00%	Second isotopic peak	Unannotated mass-shift -1.0894	Unannotated mass-shift -1.0894
1.0792	176	926	118	0	1	0	0.00%	0.11%	0.00%	Unannotated mass-shift 1.0792		
-471.2326	1	14	12	0	0	0	0.00%	0.00%	0.00%	Phosphorylation	Deletion of H	Unannotated mass-shift -414.1388
126.8988	34	13	2	0	0	0	0.00%	0.00%	0.00%	First isotopic peak	Iodination	
55.9234	107	186	4	0	0	0	0.00%	0.00%	0.00%	First isotopic peak	Unannotated mass-shift 52.9128	Second isotopic peak

-0.1182	112	490	164	0	1	0	0.00%	0.20%	0.00%	2-amino-3-oxo-butanoic_acid/Val->Pro substitution	glycosylated asparagine 18O labeling/Deamidation in presence of O18	Unannotated mass-shift -1.0894
-3.9726	50	228	52	0	0	0	0.00%	0.00%	0.00%	Addition of arginine due to transpeptidation/Addition of R	Deletion of Y	glycosylated asparagine 18O labeling/Deamidation in presence of O18
288.1568	1	21	50	0	0	0	0.00%	0.00%	0.00%	Addition of lysine due to transpeptidation/Addition of K	Acetaldehyde +26/Ala->Pro substitution	3-methyl-2-pyridyl isocyanate
27.0472	4	26	2	0	0	0	0.00%	0.00%	0.00%	Thr->Lys substitution/ethyl amino		
-341.1964	5	27	13	0	0	0	0.00%	0.00%	0.00%	Deamidation/Asn->Asp substitution/Gln->Glu substitution	Loss of arginine due to transpeptidation/Deletion of R	Deletion of W
2.089	124	682	30	0	2	0	0.00%	0.29%	0.00%	First isotopic peak	Leu/Ile->Asn substitution	Unannotated mass-shift 0.1282
-301.164	3	19	12	0	0	0	0.00%	0.00%	0.00%	Unannotated mass-shift -328.2110	Thr->Lys substitution/ethyl amino	
73.0152	8	32	23	0	1	0	0.00%	3.13%	0.00%	Carbamylation/Ala->Asn substitution	hydroxymethyl/Ala->Thr substitution/Gly->Ser substitution	
-336.1916	2	15	7	0	0	0	0.00%	0.00%	0.00%	Pyro-glu from Q/Loss of ammonia	Loss of arginine due to transpeptidation/Deletion of R	Deletion of Y
-26.0172	6	39	4	0	0	0	0.00%	0.00%	0.00%	Pro->Ala substitution		
140.0592	1	5	1	0	0	0	0.00%	0.00%	0.00%	amidination of lysines or N-terminal amines with methyl acetimidate/Ser->Gln substitution	EGS crosslinker to Lys or N-terminus following hydroxylamine cleavage	
-340.2226	5	36	11	0	0	0	0.00%	0.00%	0.00%	Acetylation/Ser->Glu substitution	Unannotated mass-shift -400.1888	Leu/Ile->Met substitution
300.1438	13	68	2	0	0	0	0.00%	0.00%	0.00%	Acetylation/Ser->Glu substitution	Carbamylation/Ala->Asn substitution	Unannotated mass-shift 215.1270
-331.157	4	12	15	0	0	0	0.00%	0.00%	0.00%	Acetylation/Ser->Glu substitution	Unannotated mass-shift -372.2198	Glu->Lys substitution
-1.9422	112	654	34	2	2	0	1.79%	0.31%	0.00%	Asp->Leu/Ile substitution		
-3.994	37	186	46	1	3	0	2.70%	1.61%	0.00%	Thr->Pro substitution		
-0.0826	205	932	229	0	7	0	0.00%	0.75%	0.00%	Lysine oxidation to aminoadipic semialdehyde	Lys->Glu substitution	
-48.1274	8	34	11	4	19	13	50.00%	55.88%	118.18%	Loss of C-terminal K from Heavy Chain of MAb/Deletion of K	Phosphorylation	
-257.1006	6	18	13	0	0	0	0.00%	0.00%	0.00%	Unannotated mass-shift -257.1376	Gln->Lys substitution	
-327.1916	3	15	7	0	0	0	0.00%	0.00%	0.00%	Acetylation/Ser->Glu substitution	Acrylamide adduct/Gly->Gln substitution/Addition of A	Unannotated mass-shift -440.2390
-269.1854	4	71	10	0	0	0	0.00%	0.00%	0.00%	Deletion of I/Deletion of L	Loss of arginine due to transpeptidation/Deletion of R	
-372.2012	7	15	5	0	1	0	0.00%	6.67%	0.00%	First isotopic peak	Amidation/Asp->Asn substitution/Glu->Gln substitution	Unannotated mass-shift -372.2198
-326.195	5	21	12	0	0	0	0.00%	0.00%	0.00%	Unannotated mass-shift -229.1430	Deletion of P	
384.2126	3	45	7	0	0	0	0.00%	0.00%	0.00%	Deamidation/Asn->Asp substitution/Gln->Glu substitution	Deletion of I/Deletion of L	Unannotated mass-shift 496.3114
1.8674	88	363	71	0	1	0	0.00%	0.28%	0.00%	Deamidation/Asn->Asp substitution/Gln->Glu substitution	Unannotated mass-shift 0.8830	
-332.1852	4	19	9	0	0	0	0.00%	0.00%	0.00%	Loss of C-terminal K from Heavy Chain of MAb/Deletion of K	Dehydration/Pyro-glu from E	Deletion of W
0.8942	158	728	192	0	4	0	0.00%	0.55%	0.00%	Pyro-glu from Q/Loss of ammonia	Lys->Gln substitution	Leu/Ile->Met substitution
171.1258	2	20	47	0	0	0	0.00%	0.00%	0.00%	Deamidation/Asn->Asp substitution/Gln->Glu substitution	Addition of lysine due to transpeptidation/Addition of K	tri-Methylation/Gly->Val substitution/Ala->Leu/Ile substitution/Propyl
112.0982	9	161	21	0	1	0	0.00%	0.62%	0.00%	Addition of lysine due to transpeptidation/Addition of K	reduction/Ser->Ala substitution/Tyr->Phe substitution	
226.1212	0	26	44	0	0	0	0.00%	0.00%	0.00%	Deamidation/Asn->Asp substitution/Gln->Glu substitution	Unannotated mass-shift 210.1616	Leu/Ile->Gln substitution/Val->Asn substitution
44.9852	27	211	25	0	1	0	0.00%	0.47%	0.00%	Oxidation to nitro		
61.9686	1	6	4	0	0	0	0.00%	0.00%	0.00%	First isotopic peak	Glu->Lys substitution	Replacement of 2 protons by zinc
-1.9584	117	697	41	1	2	0	0.86%	0.29%	0.00%	First isotopic peak	2-amino-3-oxo-butanoic_acid/Val->Pro substitution	Glu->Lys substitution
29.9998	58	376	2	0	0	0	0.00%	0.00%	0.00%	Deamidation/Asn->Asp substitution/Gln->Glu substitution	deamidation followed by esterification with ethanol	
17.9822	24	130	4	0	3	0	0.00%	2.31%	0.00%	Deamidation/Asn->Asp substitution/Gln->Glu substitution	Asn->Met substitution	
4.0092	68	1364	1	0	3	0	0.00%	0.22%	0.00%	O18 label at both C-terminal oxygens		
0.905	164	808	223	1	1	0	0.61%	0.12%	0.00%	First isotopic peak	pyrophosphorylation of Ser/Thr	Deletion of C
40.0308	6	129	3	0	1	0	0.00%	0.78%	0.00%	Propionaldehyde +40/Gly->Pro substitution		

-33.0216	4	77	12	0	0	1	0.00%	0.00%	8.33%	Pyro-glu from Q/Loss of ammonia	reduction/Ser->Ala substitution/Tyr->Phe substitution	
-72.9974	9	24	5	0	0	0	0.00%	0.00%	0.00%	First isotopic peak	amidination of lysines or N-terminal amines with methyl acetimidate/Ser->Gln substitution	Deletion of D
-301.1452	1	10	7	0	0	0	0.00%	0.00%	0.00%	Addition of lysine due to transpeptidation/Addition of K	Asn->Gly substitution/Gln->Ala substitution/Deletion of G	Unannotated mass-shift -372.2198
115.0462	8	38	10	0	0	0	0.00%	0.00%	0.00%	First isotopic peak	ubiquitinylation residue/Double Carbamidomethylation/Addition of N	
-473.2124	1	21	9	0	0	0	0.00%	0.00%	0.00%	Unannotated mass-shift -229.1430	Deletion of E	Deletion of D
-370.1852	2	21	9	0	1	0	0.00%	4.76%	0.00%	Acetylation/Ser->Glu substitution	Leu/Ile->Asp substitution	Unannotated mass-shift -414.1388
-0.9268	169	827	75	0	0	0	0.00%	0.00%	0.00%	First isotopic peak	Amidation/Asp->Asn substitution/Glu->Gln substitution	Glu->Lys substitution
294.1032	3	8	6	0	0	0	0.00%	0.00%	0.00%	Addition of M	Addition of Y	
270.1456	3	89	43	1	0	0	33.33%	0.00%	0.00%	Addition of arginine due to transpeptidation/Addition of R	ubiquitinylation residue/Double Carbamidomethylation/Addition of N	
383.2284	4	37	7	0	0	0	0.00%	0.00%	0.00%	Ubiquitination		
322.1632	1	14	25	0	0	0	0.00%	0.00%	0.00%	Addition of arginine due to transpeptidation/Addition of R	Hexose	O18 label at both C-terminal oxygens
-441.2334	3	21	4	0	0	0	0.00%	0.00%	0.00%	Deamidation/Asn->Asp substitution/Gln->Glu substitution	Thr->Val substitution	Unannotated mass-shift -440.2390
-276.1472	7	21	11	0	1	0	0.00%	4.76%	0.00%	Deletion of I/Deletion of L	Deletion of Y	
60.0028	14	42	4	0	0	0	0.00%	0.00%	0.00%	2-OH-ethyl thio-Ser/Ala->Met substitution		
355.1112	0	5	1	0	0	0	0.00%	0.00%	0.00%	dihydroxy/Pro->Glu substitution	replacement of proton with ammonium ion	Unannotated mass-shift 306.0952
82.9252	0	4	7	0	0	0	0.00%	0.00%	0.00%	Unannotated mass-shift 52.9128	hydroxymethyl/Ala->Thr substitution/Gly->Ser substitution	
1.831	47	171	40	0	0	0	0.00%	0.00%	0.00%	Lys->Glu substitution	Unannotated mass-shift 0.8830	
202.0776	6	26	15	0	0	0	0.00%	0.00%	0.00%	Addition of M	Acrylamide adduct/Gly->Gln substitution/Addition of A	
-45.0214	2	107	10	0	0	1	0.00%	0.00%	10.00%	Acetylation/Ser->Glu substitution	Deletion of S	
-0.8826	102	390	28	0	0	0	0.00%	0.00%	0.00%	First isotopic peak	2-amino-3-oxo-butanoic_acid/Val->Pro substitution	Unannotated mass-shift 0.1282
120.104	0	4	9	0	0	0	0.00%	0.00%	0.00%	Loss of C-terminal K from Heavy Chain of MAb/Deletion of K	Unannotated mass-shift 248.1986	
39.9938	21	28	11	0	2	0	0.00%	7.14%	0.00%	S-carbamoylmethylcysteine cyclization (N-terminus)/Glyoxal-derived hydroimiadazolone		
-344.1698	5	31	11	0	0	0	0.00%	0.00%	0.00%	Unannotated mass-shift -229.1430	Deletion of D	
346.04	0	288	1	0	0	0	0.00%	0.00%	0.00%	Unannotated mass-shift 284.1268	Replacement of 2 protons by zinc	
-147.0354	2	57	16	0	0	0	0.00%	0.00%	0.00%	reduction/Ser->Ala substitution/Tyr->Phe substitution	Removal of initiator methionine from protein N-terminus/Deletion of M	

Appendix D

MSFRAGGER MANUAL

Introduction

MSFragger is an ultrafast database search tool for peptide identifications in mass spectrometry-based proteomics. It differs from conventional search engines by computing similarity scores in a fragment-centric fashion using a theoretical fragment index of candidate peptides. The speed of MSFragger makes it particularly suitable for ‘open’ database searches, where the precursor mass tolerance is set to hundreds of Daltons, for the identification of modified peptides. MSFragger is implemented in the cross-platform Java programming language and is compatible with standard proteomics file formats such as MGF/mzXML/mzML/pepXML.

Equipment

Computer Hardware requirements

The processor requirements of MSFragger depends on the complexity of your search (and your patience to wait for search results). For an open search (500Da precursor mass window) using a tryptic digest of the human proteome, a single processor core can search roughly 40,000 MS/MS spectra in under an hour. MSFragger scales well with the number of processor cores and runtimes of under 2 minutes per file have been achieved using a 28-core workstation. A desktop workstation with a quad core processor is sufficient for most simple workflows.

MSFragger requires substantial amounts of memory due to its in-memory fragment index. While MSFragger can operate with less memory than needed to store the fragment index, it will cause index fragmentation where it breaks the search into multiple passes, searching each input file against a small segment of the index at a time (which greatly increases the runtime). For the human Uniprot protein database with reversed decoys, approximately 3700 MB of memory is needed to prevent index fragmentation. The actual size of the fragment index is substantially lower (MSFragger uses a very conservative estimate of the available free memory to avoid out of memory situations). Specifying common modifications may boost memory requirements to 6 GB. Semi-tryptic, non-enzymatic, and phospho searches may take tens of gigabytes to avoid fragmented searches. Limiting the range of peptide lengths can reduce the search space and reduce memory consumption in such cases. While fragment index fragmentation is undesirable, it may be unavoidable in certain instances.

We recommend at least 8GB of memory for workflows involving standard tryptic digestions.

Operating System requirements

MSFragger has been tested on Mac OS X, Windows 7, and a number of Linux distributions. Note that a 64-bit operating system is required to access more than 4GB of memory.

Java requirements

MSFragger is written using Java 1.8 and requires the Java 8 Runtime Environment. We recommend the Oracle Java 8 Runtime (download and installation instructions are available at www.java.com).

Procedure

Preparing Input Files

Mass spectrometry data must first be converted to one of the supported MS/MS input formats of MGF, mzXML, or mzML. A popular option for converting from vendor file inputs and between various input formats is Proteowizard (proteowizard.sourceforge.net). MSFragger determines the appropriate data parser to use based on the file extension (.mgf for MGF, .mzXML for mzXML, and .mzML for mzML) and does not make inferences from file contents (i.e. naming a mzML file with the .mzXML extension will lead to unpredictable results or crashes).

The protein database must be supplied in FASTA format. MSFragger does not have the capability to generate decoys internally so they must be generated externally and appended to the protein database before running MSFragger.

Configuring MSFragger

Extract the MSFragger.jar into your working directory along with the sample configuration file called fragger.params. MSFragger is configured using a text parameters file. The parameters file is passed as the first argument to MSFragger and has no restrictions on names or file extensions (so one might want to name their configuration files to be more descriptive such as Uniprot_open_withmods.txt) after editing the parameters file for a particular analysis.

Parameter names are given left of the equal sign and parameter values are given to the right (e.g. num_threads = 4). White spaces are trimmed from the ends of each value by MSFragger. All text to the right of (and including) the # sign of each line is discarded so # can be used for comments in the parameters file.

Table D-1 Listing of MSFragger search parameters.

General Parameters

num_threads	Number of CPU threads to use, should be set to the number of logical processors; a value of 0 (auto-detect) will cause MSFragger to use the auto-detected number of processors Default: 0
database_name	Path to the protein database file in FASTA format

Search Tolerances

precursor_mass_tolerance	Precursor mass tolerance (window is +/- this value) Default: 20
precursor_mass_units	Precursor mass tolerance units (0 for Da, 1 for ppm) Default: 1
precursor_true_tolerance	True precursor mass tolerance (window is +/- this value). Used for tie breaker of results (in spectrally ambiguous cases) and zero bin boosting in open searches (0 disables these features). This option is STRONGLY recommended for open searches. Default: 0
precursor_true_units	True precursor mass tolerance units (0 for Da, 1 for ppm) Default: 1
fragment_mass_tolerance	Fragment mass tolerance (window is +/- this value) Default: 20
fragment_mass_units	Fragment mass tolerance units (0 for Da, 1 for ppm) Default: 1

isotope_error	Isotope correction for MS/MS events triggered on isotopic peaks. Should be set to 0 (disabled) for open search or 0/1/2 for correction of narrow window searches. Shifts the precursor mass window to multiples of this value multiplied by the mass of C13-C12. Default: 0

In-silico Digestion Parameters

search_enzyme_name	Name of enzyme to be written to the pepXML file. Default: Trypsin
search_enzyme_cutafter	Residues after which the enzyme cuts (specified as a string of amino acids) Default: KR
search_enzyme_butnotafter	Residues that the enzyme will not cut before (misnomer: should really be called butnotbefore) Default: P
num_enzyme_termini	Number of enzyme termini (0, 1, or 2 for non-enzymatic, semi-enzymatic, fully-enzymatic) Default: 2
allowed_missed_cleavage	Allowed number of missed cleavages Default: 2
digest_min_length	Minimum length of peptides to be generated during in-silico digestion Default: 7
digest_max_length	Maximum length of peptides to be generated during in-silico digestion Default: 64
digest_mass_range	Mass range of peptides to be generated during in-silico digestion in Daltons (specified as a space separated range) Default: 500.0 5000.0

Variable Modification Parameters

clip_nTerm_M	Specifies the trimming of a protein N-terminal methionine as a variable modification (0 or 1) Default: 0

variable_mod_01 .. 07	Sets variable modifications. (variable_mod_01 to variable_mod_07). Space separated values with 1st value being the modification mass and the second being the residues (specified consecutively as a string) it modifies. * is used to represent any amino acid [is a modifier for protein N-terminal] is a modifier for protein C-terminal n is a modifier for peptide N-terminal c is a modifier for peptide C-terminal Syntax Examples: 15.9949 M (for oxidation on methionine) 79.66331 STY (for phosphorylation) -17.0265 nQnC (for pyro-Glu or loss of ammonia at peptide N-terminal) Example (M oxidation and N-terminal acetylation): variable_mod_01 = 15.9949 M variable_mod_02 = 42.0106 [*
allow_multiple_variable_mods_on_residue	Allow each amino acid to be modified by multiple variable modifications (0 or 1) Default: 1
max_variable_mods_per_mod	Maximum number of residues that can be occupied by each variable modification (maximum of 5). Default: 2
max_variable_mods_combinations	Maximum allowed number of modified variably modified peptides from each peptide sequence, (maximum of 65534). If a greater number than the maximum is generated, only the unmodified peptide is considered. Default: 5000

Spectrum Processing Parameters

minimum_peaks	Minimum number of peaks in experimental spectrum for matching Default: 10
use_topN_peaks	Pre-process experimental spectrum to only use top N peaks Default: 50
minimum_ratio	Filters out all peaks in experimental spectrum less intense than this multiple of the base peak intensity Default: 0.0

clear_mz_range	Removes peaks in this m/z range prior to matching. Useful for iTRAQ/TMT experiments (i.e. 0.0 150.0). Default: 0.0 0.0
max_fragment_charge	Maximum charge state for theoretical fragments to match (1-4). Default: 2
override_charge	Ignores precursor charge and uses charge state specified in precursor_charge range (0 or 1) Default: 0
precursor_charge	Assume range of potential precursor charge states. Only relevant when override_charge is set to 1. Specified as space separated range of integers. Default: 1 4

Open Search Features

track_zero_topN	Track top N unmodified peptide results separately from main results internally for boosting features. Should be set to a number greater than output_report_topN if zero bin boosting is desired. Default: 0
zero_bin_accept_expect	Ranks a zero-bin hit above all non-zero-bin hit if it has expectation less than this value. Default: 0.0
zero_bin_mult_expect	Multiplies expect value of PSMs in the zero-bin during results ordering (set to less than 1 for boosting). Default: 1.0
add_topN_complementary	Inserts complementary ions corresponding to the top N most intense fragments in each experimental spectra. Useful for recovery of modified peptides near C-terminal in open search. Should be set to 0 (disabled) otherwise. Default: 0

Modeling and Output Parameters

min_fragments_modelling	Minimum number of matched peaks in PSM for inclusion in statistical modeling Default: 3

min_matched_fragments	Minimum number of matched peaks for PSM to be reported. We recommend a minimum of 4 for narrow window searching and 6 for open searches. Default: 4
output_file_extension	File extension of output files Default: pep.xml
output_format	File format of output files (pepXML or tsv) Default: pepXML
output_report_topN	Reports top N PSMs per input spectrum Default: 1
output_max_expect	Suppresses reporting of PSM if top hit has expectation greater than this threshold Default: 50.0

Static Modification Parameters

add_Cterm_peptide	Statically add mass in Da to C-terminal of peptide Default: 0.0
add_Nterm_peptide	Statically add mass in Da to N-terminal of peptide Default: 0.0
add_Cterm_protein	Statically add mass in Da to C-terminal of protein Default: 0.0
add_Nterm_protein	Statically add mass in Da to N-terminal of protein Default: 0.0
add_C_cysteine ... add_X_usertext	Statically add mass to cysteine (or whatever amino acid is specified after ‘add_’). Examples: add_C_cysteine = 57.021464 add_K_lysine = 144.1021 Default: 0.0

Running MSFragger

Performance Considerations for Batch Processing

MSFragger allows multiple MS/MS input files to be processed in a batch. Passing multiple files to MSFragger at once allows MSFragger to reuse the fragment index for subsequent MS/MS run. This is particularly important for narrow window searches which may only take fractions of a second.

On computers or compute clusters with many processor cores, we highly recommended that MSFragger is set to process files sequentially with all available processor cores rather than running multiple instances of MSFragger in parallel (assigning a smaller number of cores to each). This reduces initialization times and allows the fragment index to be re-used, at the same time reducing overall memory requirements.

Launching MSFragger

Ensure that you have placed MSFragger.jar in your working directory and have modified the parameters file to reference your protein database. MSFragger generates auxiliary files during database search so it is critical that **MSFragger must have write access to the directories containing the protein database AND the MS/MS data files**.

Determine the amount of system memory available that you would like to make available to MSFragger. This will be specified by the Java maximum heap size parameter -Xmx (e.g. -Xmx3700M for 3700 MB or -Xmx8G for 8GB).

MSFragger takes the first argument as the input parameters file, followed by a list of one or more MS/MS data files.

Examples:

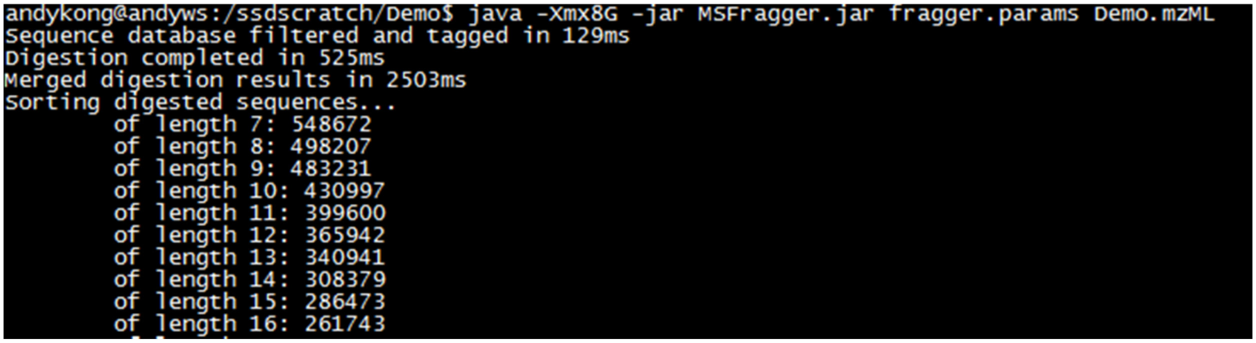
```
java -Xmx8G -jar MSFragger.jar fragger.params HeLa_run1.mzML HeLa_run2.mzML
java -Xmx8G -jar MSFragger.jar fragger.params *.mzML
```

The **-Xmx flag is very important** to ensure that MSFragger has access to sufficient memory to efficiently perform the search as the default max heap setting in Java is ¼ of total system memory (which is insufficient for optimal performance). We recommend that you can allocate a minimum of 4G or 6G for standard tryptic digestions.

Expected Behavior

The first time running MSFragger on a new protein database or set of search parameters with a given database, it will first perform an in-silico digestion, create, and cache the peptide index (in .pepindex files adjacent to where the FASTA database is stored). These pepindex files can be safely removed at any time and should be removed to free up disk space when a set of search parameters is no longer used (MSFragger will automatically re-generate the index as needed).

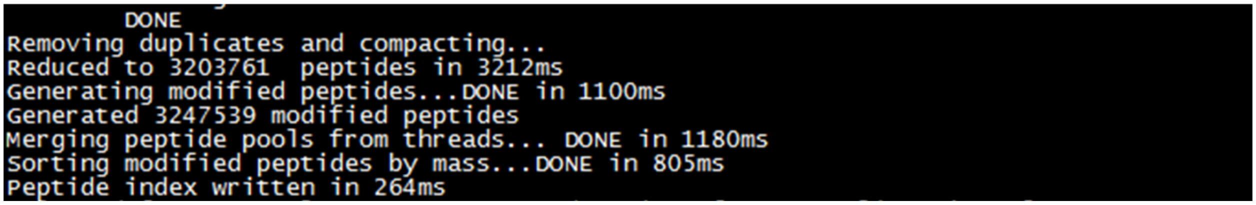
The process begins with filtering and in-silico digestion subject to the digestion parameters.



```
andykong@andyws:/ssdscratch/Demo$ java -Xmx8G -jar MSFragger.jar fragger.params Demo.mzML
Sequence database filtered and tagged in 129ms
Digestion completed in 525ms
Merged digestion results in 2503ms
Sorting digested sequences...
        of length 7: 548672
        of length 8: 498207
        of length 9: 483231
        of length 10: 430997
        of length 11: 399600
        of length 12: 365942
        of length 13: 340941
        of length 14: 308379
        of length 15: 286473
        of length 16: 261743
```

Figure D-1 In-silico digestion in MSFragger.

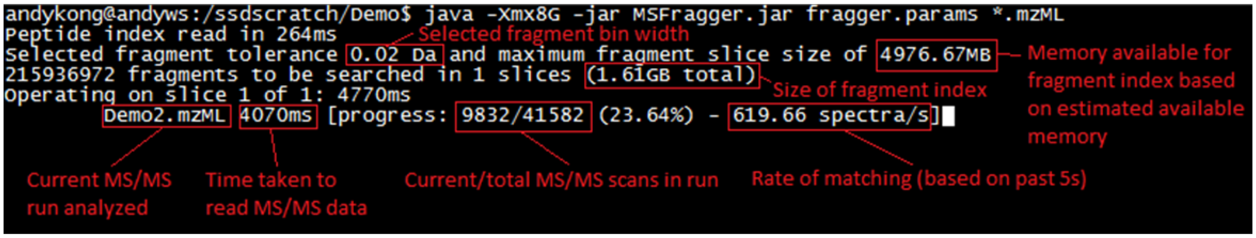
Followed by peptide sorting and de-duplication. The non-redundant set of peptides are then evaluated to generate the set of variably modified peptides (based on the specified variable modifications) which are then sorted by mass and stored.



```
        DONE
Removing duplicates and compacting...
Reduced to 3203761  peptides in 3212ms
Generating modified peptides...DONE in 1100ms
Generated 3247539 modified peptides
Merging peptide pools from threads... DONE in 1180ms
Sorting modified peptides by mass...DONE in 805ms
Peptide index written in 264ms
```

Figure D-2 Peptide index generation in MSFragger.

After peptide index generation is complete (or is read from disk in the below screenshot). MSFragger selects the fragment index bin width to use and estimates the memory available for fragment index storage based on the available memory (in this case, 8GB of memory was made available to the Java Virtual Machine, of which MSFragger estimates that 4976.67MB can be safely reserved for fragment index operations). It then computes the number of theoretical fragments to be generated for the entire index, the number of slices or iterations (in multi-pass searches when there is insufficient memory), and the total amount of memory represented by the entire fragment index. The fragment index is then generated, and a time is reported for the index generation time (at the end of each Operating on slice 1 of X: line, 4770 ms below). If the maximum fragment slice size is very small compared to your desired amount of system memory or the number of slices is unexpectedly high, double check that the -Xmx flag is correctly set. Search begins and the current file is reported, along with the time needed to read and pre-process the MS/MS data, along with current search progress.

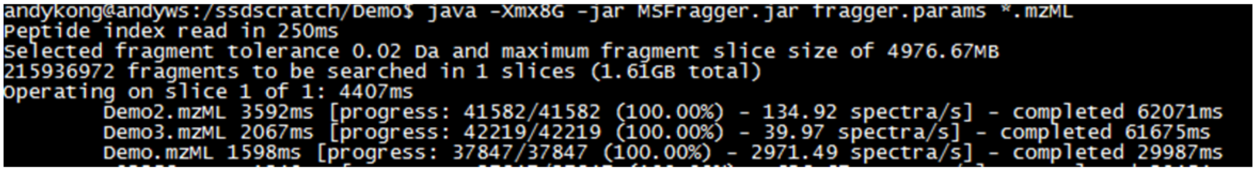


```
andykong@andyws:/ssdscratch/Demo$ java -Xmx8G -jar MSFragger.jar fragger.params *.mzML
Peptide index read in 264ms
Selected fragment tolerance 0.02 Da and maximum fragment slice size of 4976.67MB
215936972 fragments to be searched in 1 slices (1.61GB total)
Operating on slice 1 of 1: 4770ms
        Demo2.mzML 4070ms [progress: 9832/41582 (23.64%) - 619.66 spectra/s]
```

Selected fragment bin width
Memory available for fragment index based on estimated available memory
Size of fragment index
Current MS/MS run analyzed
Time taken to read MS/MS data
Current/total MS/MS scans in run
Rate of matching (based on past 5s)

Figure D-3 Fragment index searching in MSFragger.

At the completion of the search, a completed time is reported, and the results are written to disk in the same folder as the MS/MS data (if they are not in the same folder as your working directory). Note that there is a current bug that causes MSFragger to incorrectly display the average rate of matching at the conclusion of the run (although the total time can be divided by the total number of spectra to calculate this value).



```
andykong@andyws:/ssdscratch/Demo$ java -Xmx8G -jar MSFragger.jar fragger.params *.mzML
Peptide index read in 250ms
Selected fragment tolerance 0.02 Da and maximum fragment slice size of 4976.67MB
215936972 fragments to be searched in 1 slices (1.61GB total)
Operating on slice 1 of 1: 4407ms
        Demo2.mzML 3592ms [progress: 41582/41582 (100.00%) - 134.92 spectra/s] - completed 62071ms
        Demo3.mzML 2067ms [progress: 42219/42219 (100.00%) - 39.97 spectra/s] - completed 61675ms
        Demo.mzML 1598ms [progress: 37847/37847 (100.00%) - 2971.49 spectra/s] - completed 29987ms
```

Figure D-4 MSFragger searching in batch mode.

Output Files

Table D-2 Listing of MSFragger output files.

.fragtmp	In cases of fragment index fragmentation (in limited memory scenarios), MSFragger will iteratively load each MS/MS run and search loaded spectra against the current index slice before working on the next index slice. The partial search results are then stored in these .fragtmp files. In the event that MSFragger is terminated in the middle of a search, it will recover its partial results using these files. At the end of the last index slice, MSFragger will read all such .fragtmp files and generate an aggregated results file (identical to one that would be generated if it had the memory to search against all peptides in a single pass). These .fragtmp files are then automatically deleted. These can be safely removed if you no longer wish to continue an aborted search or if MSFragger somehow fails to remove them at the conclusion of a successful search. Location: Same directory as MS/MS files
.pepindex	MSFragger stores the computed peptide index in .pepindex files adjacent to the protein database files to remove the need to re-compute the index if search parameters are unchanged in subsequent runs. These .pepindex indices can be safely removed and MSFragger will re-compute the index again at runtime if needed. Location: Same directory as protein database

Results Files (eg. .pep.xml)	These are the pepXML or TSV output files containing the peptide identifications. The file extension is specified in the search parameters so specifying a .pep.xml extension with output_format = tsv will output .pep.xml files with TSV content. Location: Same directory as MS/MS files

Interpretation of Output

For pepXML outputs, these can be used for downstream processing using PeptideProphet in TPP directly. For viewing of results or conversion to other peptide identification result formats for use in other pipelines or tools that do not support pepXML, we recommend first converting to the mzIdentML format using the tool idconvert as part of the ProteoWizard package. The pepXML generated by MSFragger validates against v 1.18 of the pepXML schema and should be compatible with any downstream tools supporting the pepXML format.

The output fields of the TSV file produced by MSFragger are listed below:

ScanID
Precursor neutral mass (Da)
Retention time (minutes)
Precursor charge
Hit rank
Peptide Sequence
Upstream Amino Acid
Downstream Amino Acid
Protein
Matched fragment ions
Total possible number of matched theoretical fragment ions
Neutral mass of peptide (including any variable modifications) (Da)
Mass difference
Number of tryptic termini
Number of missed cleavages
Variable modifications detected
(starts with M, separated by |, formated as position,mass)
Hyperscore
Next score
Intercept of expectation model (expectation in log space)
Slope of expectation model (expectation in log space)

REFERENCES

[1] E.S. Lander, L.M. Linton, B. Birren, C. Nusbaum, M.C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J.P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J.C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R.H. Waterston, R.K. Wilson, L.W. Hillier, J.D. McPherson, M.A. Marra, E.R. Mardis, L.A. Fulton, A.T. Chinwalla, K.H. Pepin, W.R. Gish, S.L. Chissoe, M.C. Wendl, K.D. Delehaunty, T.L. Miner, A. Delehaunty, J.B. Kramer, L.L. Cook, R.S. Fulton, D.L. Johnson, P.J. Minx, S.W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J.F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R.A. Gibbs, D.M. Muzny, S.E. Scherer, J.B. Bouck, E.J. Sodergren, K.C. Worley, C.M. Rives, J.H. Gorrell, M.L. Metzker, S.L. Naylor, R.S. Kucherlapati, D.L. Nelson, G.M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H.M. Yang, J. Yu, J. Wang, G.Y. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S.Z. Qin, R.W. Davis, N.A. Federspiel, A.P. Abola, M.J. Proctor, R.M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D.R. Cox, M. V Olson, R. Kaul, N. Shimizu, K. Kawasaki, S. Minoshima, G.A. Evans, M. Athanasiou, R. Schultz, B.A. Roe, F. Chen, H.Q. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W.R. McCombie, M. de la Bastide, N. Dedhia, H. Blocker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J.A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D.G. Brown, C.B. Burge, L. Cerutti, H.C. Chen, D. Church, M. Clamp, R.R. Copley, T. Doerks, S.R. Eddy, E.E. Eichler, T.S. Furey, J. Galagan, J.G.R. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W.H. Jang, L.S. Johnson, T.A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W.J. Kent, P. Kitts, E. V Koonin, I. Korf, D. Kulp, D. Lancet, T.M. Lowe, A. McLysaght, T. Mikkelsen, J. V Moran, N. Mulder, V.J. Pollara, C.P. Ponting, G. Schuler, J.R. Schultz, G. Slater, A.F.A. Smit, E. Stupka, J. Szustakowki, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y.I. Wolf, K.H. Wolfe, S.P. Yang, R.F. Yeh, F. Collins, M.S. Guyer, J. Peterson, A. Felsenfeld, K.A. Wetterstrand, A. Patrinos, M.J. Morgan, Initial sequencing and analysis of the human genome, Nat. 409 (2001) 860–921. doi:10.1038/35057062.

[2] J.C. Venter, M.D. Adams, E.W. Myers, P.W. Li, R.J. Mural, G.G. Sutton, H.O. Smith, M.

Yandell, C.A. Evans, R.A. Holt, J.D. Gocayne, P. Amanatides, R.M. Ballew, D.H. Huson, J.R. Wortman, Q. Zhang, C.D. Kodira, X.H. Zheng, L. Chen, M. Skupski, G. Subramanian, P.D. Thomas, J. Zhang, G.L. Gabor Miklos, C. Nelson, S. Broder, A.G. Clark, J. Nadeau, V.A. McKusick, N. Zinder, A.J. Levine, R.J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, F. Di V, P. Dunn, K. Eilbeck, C. Evangelista, A.E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T.J. Heiman, M.E. Higgins, R.R. Ji, Z. Ke, K.A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V Merkulov, N. Milshina, H.M. Moore, A.K. Naik, V.A. Narayan, B. Neelam, D. Nusskern, D.B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Wang, A. Wang, X. Wang, J. Wang, M. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M.L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y.H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N.N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J.F. Abril, R. Guigo, M.J. Campbell, K. V Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y.H. Chiang, M. Coyne, C. Dahlke, A. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, The sequence of the human genome, Sci. (Washingt. D C). 291 (2001) 1304–1351. doi:10.1126/science.1058040.

[3] M.-S. Kim, S. Pinto, D. Getnet, R. Nirujogi, S. Manda, R. Chaerkady, A. Madugundu, D. Kelkar, R. Isserlin, S. Jain, J. Thomas, B. Muthusamy, L.-R. Pamela, P. Kumar, N. Sahasrabuddhe, L. Balakrishnan, J. Advani, B. George, S. Renuse, L. Selvan, A. Patil, V. Nanjappa, A. Radhakrishnan, S. Prasad, T. Subbannayya, R. Raju, M. Kumar, S. Sreenivasamurthy, A. Marimuthu, G. Sathe, S. Chavan, K. Datta, Y. Subbannayya, A. Sahu, S. Yelamanchi, S. Jayaram, P. Rajagopalan, J. Sharma, K. Murthy, N. Syed, R. Goel, A. Khan, S. Ahmad, G. Dey, K. Mudgal, A. Chatterjee, T.-C. Huang, J. Zhong, X. Wu, P. Shaw, D. Freed, M. Zahari, K. Mukherjee, S. Shankar, A. Mahadevan, H. Lam, C. Mitchell, S. Shankar, P. Satishchandra, J. Schroeder, R. Sirdeshmukh, A. Maitra, S. Leach, C. Drake, M. Halushka, T. Prasad, R. Hruban, C. Kerr, G. Bader, I.-D. Christine, H. Gowda, A. Pandey, A draft map of the human proteome, Nature. 509 (2014) 575–581.

doi:10.1038/nature13302.

[4] M. Wilhelm, J. Schlegl, H. Hahne, A. Moghaddas Gholami, M. Lieberenz, M.M. Savitski, E. Ziegler, L. Butzmann, S. Gessulat, H. Marx, T. Mathieson, S. Lemeer, K. Schnatbaum, U. Reimer, H. Wenschuh, M. Mollenhauer, J. Slotta-Huspenina, J.-H. Boese, M. Bantscheff, A. Gerstmair, F. Faerber, B. Kuster, Mass-spectrometry-based draft of the human proteome., Nature. 509 (2014) 582–7. doi:10.1038/nature13319.

[5] G.S. Omenn, L. Lane, E.K. Lundberg, R.C. Beavis, A.I. Nesvizhskii, E.W. Deutsch, Metrics for the human proteome project 2015: Progress on the human proteome and guidelines for high-confidence protein identification, J. Proteome Res. 14 (2015) 3452–3460. doi:10.1021/acs.jproteome.5b00499.

[6] I. Ezkurdia, J. V??zquez, A. Valencia, M. Tress, Analyzing the first drafts of the human proteome, J. Proteome Res. 13 (2014) 3854–3855. doi:10.1021/pr500572z.

[7] A.I. Nesvizhskii, Proteogenomics: concepts, applications and computational strategies, Nat. Methods. 11 (2014) 1114–1125. doi:10.1038/nmeth.3144.

[8] M.M. Savitski, M. WIlhelm, H. Hahne, B. Kuster, M. Bantscheff, A scalable approach for protein false discovery rate estimation in large proteomic data sets, Mol. Cell. Proteomics. 14 (2015) mcp.M114.046995. doi:10.1074/mcp.M114.046995.

[9] S.D. Patterson, R.H. Aebersold, Proteomics: the first decade and beyond., Nat. Genet. 33 Suppl (2003) 311–23. doi:10.1038/ng1106.

[10] B. Boeckmann, A. Bairoch, R. Apweiler, M.C. Blatter, A. Estreicher, E. Gasteiger, M.J. Martin, K. Michoud, C. O’Donovan, I. Phan, S. Pilbout, M. Schneider, The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003, Nucleic Acids Res. 31 (2003) 365–370. doi:10.1093/nar/gkg095.

[11] T. Hubbard, D. Barker, E. Birney, G. Cameron, Y. Chen, L. Clark, T. Cox, J. Cuff, V. Curwen, T. Down, R. Durbin, E. Eyras, J. Gilbert, M. Hammond, L. Huminiecki, A. Kasprzyk, H. Lehvaslaiho, P. Lijnzaad, C. Melsopp, E. Mongin, R. Pettett, M. Pocock, S. Potter, A. Rust, E. Schmidt, S. Searle, G. Slater, J. Smith, W. Spooner, A. Stabenau, J. Stalker, E. Stupka, A. Ureta-Vidal, I. Vastrik, M. Clamp, The Ensembl genome database project, Nucleic Acids Res. 30 (2002) 38–41. doi:10.1093/NAR/30.1.38.

[12] X. Wang, B. Zhang, J. Wren, CustomProDB: An R package to generate customized protein databases from RNA-Seq data for proteomics search, Bioinformatics. 29 (2013) 3235–3237. doi:10.1093/bioinformatics/btt543.

[13] B. Zhang, J. Wang, X. Wang, J. Zhu, Q. Liu, Z. Shi, M.C. Chambers, L.J. Zimmerman, K.F. Shaddox, S. Kim, S.R. Davies, S. Wang, P. Wang, C.R. Kinsinger, R.C. Rivers, H. Rodriguez, R.R. Townsend, M.J.C. Ellis, S.A. Carr, D.L. Tabb, R.J. Coffey, R.J.C. Slebos, D.C. Liebler, Proteogenomic characterization of human colon and rectal cancer., Nature. 513 (2014) 382–7. doi:10.1038/nature13438.

[14] B. Bánfai, H. Jia, J. Khatun, E. Wood, B. Risk, W.E. Gundling, A. Kundaje, H.P.

Gunawardena, Y. Yu, L. Xie, K. Krajewski, B.D. Strahl, X. Chen, P. Bickel, M.C. Giddings, J.B. Brown, L. Lipovich, Long noncoding RNAs are rarely translated in two human cell lines, Genome Res. 22 (2012) 1646–1657. doi:10.1101/gr.134767.111.

[15] Y.-K. Paik, S.-K. Jeong, G.S. Omenn, M. Uhlen, S. Hanash, S.Y. Cho, H.-J. Lee, K. Na, E.-Y. Choi, F. Yan, F. Zhang, Y. Zhang, M. Snyder, Y. Cheng, R. Chen, G. Marko-Varga, E.W. Deutsch, H. Kim, J.-Y. Kwon, R. Aebersold, A. Bairoch, A.D. Taylor, K.Y. Kim, E.-Y. Lee, D. Hochstrasser, P. Legrain, W.S. Hancock, The Chromosome-Centric Human Proteome Project for cataloging proteins encoded in the genome, Nat. Biotechnol. 30 (2012) 221–223. doi:10.1038/nbt.2152.

[16] M. Mann, O.N. Jensen, Proteomic analysis of post-translational modifications., Nat. Biotechnol. 21 (2003) 255–61. doi:10.1038/nbt0303-255.

[17] J.R. Yates, The Revolution and Evolution of Shotgun Proteomics for Large-Scale Proteome Analysis The Revolution and Evolution of Shotgun Proteomics for Large-Scale Proteome Analysis, J. Am. Chem. Soc. 135 (2013) 1629–1640.

[18] D. Arnott, J. Shabanowitz, D.F. Hunt, Mass spectrometry of proteins and peptides: Sensitive and accurate mass measurement and sequence analysis, Clin. Chem. 39 (1993) 2005–2010.

[19] D.C. Stahl, K.M. Swiderek, M.T. Davis, T.D. Lee, Data-Controlled Automation of Liquid Chromatography / Tandem Mass Spectrometry Analysis of Peptide Mixtures, Am. Soc. Mass Spectrom. 7 (1996) 532–540. doi:10.1016/1044-0305(96)00057-8.

[20] A.I. Nesvizhskii, A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics, J. Proteomics. 73 (2010) 2092–2123. doi:10.1016/j.jprot.2010.08.009.

[21] C. Spahr, M. Davis, M.D. McGinley, J.H. Robinson, E.J. Bures, J. Beierle, J. Mort, P.L. Courchesne, K. Chen, R.C. Wahl, W. Yu, R. Luethy, S.D. Patterson, Towards defining the urinary proteome using liquid chromatography-tandem mass spectrometry I. Profiling an unfractionated tryptic digest, Proteomics. 1 (2001) 93–107. doi:10.1002/1615-9861(200101)1:1<93::AID-PROT93>3.0.CO;2-3.

[22] T. Geiger, A. Wehner, C. Schaab, J. Cox, M. Mann, Comparative Proteomic Analysis of Eleven Common Cell Lines Reveals Ubiquitous but Varying Expression of Most Proteins, Mol. Cell. Proteomics. 11 (2012) M111.014050-M111.014050. doi:10.1074/mcp.M111.014050.

[23] N.A. Kulak, P.E. Geyer, M. Mann, Loss-less nano-fractionator for high sensitivity , high coverage proteomics, (2017) 1–24. doi:10.1074/mcp.O116.065136.

[24] A.S. Hebert, A.L. Richards, D.J. Bailey, A. Ulbrich, E.E. Coughlin, M.S. Westphall, J.J. Coon, The One Hour Yeast Proteome, Mol. Cell. Proteomics. 13 (2014) 339–347. doi:10.1074/mcp.M113.034769.

[25] L.M.F. de Godoy, J. V Olsen, J. Cox, M.L. Nielsen, N.C. Hubner, F. Fröhlich, T.C.

Walther, M. Mann, Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast., Nature. 455 (2008) 1251–1254. doi:10.1038/nature07341.

[26] G.L. Andrews, B.L. Simons, J.B. Young, A.M. Hawkridge, D.C. Muddiman, Performance characteristics of a new hybrid quadrupole time-of-flight tandem mass spectrometer (TripleTOF 5600), Anal Chem. 83 (2011) 5442–5446. doi:10.1021/ac200812d.

[27] A. Michalski, E. Damoc, J.-P. Hauschild, O. Lange, A. Wieghaus, A. Makarov, N. Nagaraj, J. Cox, M. Mann, S. Horning, Mass Spectrometry-based Proteomics Using Q Exactive, a High-performance Benchtop Quadrupole Orbitrap Mass Spectrometer., Mol. Cell. Proteomics. 10 (2011) M111.011015. doi:10.1074/mcp.M111.011015.

[28] P. Jones, R.G. Côté, L. Martens, A.F. Quinn, C.F. Taylor, W. Derache, H. Hermjakob, R. Apweiler, PRIDE: a public repository of protein and peptide identifications for the proteomics community., Nucleic Acids Res. 34 (2006) D659-63. doi:10.1093/nar/gkj138.

[29] J. V Olsen, J. Schwartz, J. Griep-Raming, M.L. Nielsen, E. Damoc, E. Denisov, O. Lange, P. Remes, D. Taylor, M. Splendore, E.R. Wouters, M. Senko, A. Makarov, M. Mann, S. Horning, A dual pressure linear ion trap Orbitrap instrument with very high sequencing speed., Mol. Cell. Proteomics. 8 (2009) 2759–2769. doi:10.1074/mcp.M900375-MCP200.

[30] M. Scigelova, A. Makarov, Orbitrap mass analyzer - Overview and applications in proteomics, Proteomics. 1 (2006) 16–21. doi:10.1002/pmic.200600528.

[31] a. Michalski, E. Damoc, O. Lange, E. Denisov, D. Nolting, M. Muller, R. Viner, J. Schwartz, P. Remes, M. Belford, J.-J. Dunyach, J. Cox, S. Horning, M. Mann, a. Makarov, Ultra High Resolution Linear Ion Trap Orbitrap Mass Spectrometer (Orbitrap Elite) Facilitates Top Down LC MS/MS and Versatile Peptide Fragmentation Modes, Mol. Cell. Proteomics. 11 (2012) O111.013698-O111.013698. doi:10.1074/mcp.O111.013698.

[32] Z. Zhang, Prediction of Low-Energy Collision-Induced Dissociation Spectra of Peptides with Three or More Charges, Anal. Chem. 77 (2005) 6364–6373. doi:10.1021/ac050857k.

[33] J.K. Eng, A.L. Mccormack, J.R. Yates, An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database, Am. Soc. Mass Spectrom. 5 (1994) 976–989. doi:10.1016/1044-0305(94)80016-2.

[34] J.K. Eng, T.A. Jahan, M.R. Hoopmann, Comet: An open-source MS/MS sequence database search tool, Proteomics. 13 (2013) 22–24. doi:10.1002/pmic.201200439.

[35] D. Creasy, Perkins DN , Pappin DJ , Creasy DM , Cottrell JS .. Probability-based protein identification by searching sequence databases using mass spectrometry data . Electrophoresis 20 : 3551-, Electrophoresis. 20 (1999) 3551–3567. doi:10.1002/(SICI)1522-2683(19991201)20.

[36] R. Craig, R.C. Beavis, TANDEM: Matching proteins with tandem mass spectra, Bioinformatics. 20 (2004) 1466–1467. doi:10.1093/bioinformatics/bth092.

[37] J. Cox, N. Neuhauser, A. Michalski, R.A. Scheltema, J. V. Olsen, M. Mann, Andromeda:

A peptide search engine integrated into the MaxQuant environment, J. Proteome Res. 10 (2011) 1794–1805. doi:10.1021/pr101065j.

[38] S. Kim, P.A. Pevzner, MS-GF+ makes progress towards a universal database search tool for proteomics, Nat. Commun. 5 (2014) 5277. doi:10.1038/ncomms6277.

[39] H. Lam, E.W. Deutsch, J.S. Eddes, J.K. Eng, N. King, S.E. Stein, R. Aebersold, Development and validation of a spectral library searching method for peptide identification from MS/MS, Proteomics. 7 (2007) 655–667. doi:10.1002/pmic.200600625.

[40] S. Tanner, H. Shu, A. Frank, L.C. Wang, E. Zandi, M. Mumby, P.A. Pevzner, V. Bafna, InsPecT: Identification of posttranslationally modified peptides from tandem mass spectra, Anal. Chem. 77 (2005) 4626–4639. doi:10.1021/ac050102d.

[41] S. Na, N. Bandeira, E. Paek, Fast Multi-blind Modification Search through Tandem Mass Spectrometry, Mol. Cell. Proteomics. 11 (2012) M111.010199-M111.010199. doi:10.1074/mcp.M111.010199.

[42] A. Frank, P. Pevzner, PepNovo: De novo peptide sequencing via probabilistic network modeling, Anal. Chem. 77 (2005) 964–973. doi:10.1021/ac048788h.

[43] B. Ma, K. Zhang, C. Hendrie, C. Liang, M. Li, A. Doherty-Kirby, G. Lajoie, PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry., Rapid Commun. Mass Spectrom. 17 (2003) 2337–2342. doi:10.1002/rcm.1196.

[44] X. Han, L. He, L. Xin, B. Shan, B. Ma, PeaksPTM: Mass spectrometry-based identification of peptides with unspecified modifications, J. Proteome Res. 10 (2011) 2930–2936. doi:10.1021/pr200153k.

[45] J. Zhang, L. Xin, B. Shan, W. Chen, M. Xie, D. Yuen, W. Zhang, Z. Zhang, G. a. Lajoie, B. Ma, PEAKS DB: De Novo Sequencing Assisted Database Search for Sensitive and Accurate Peptide Identification, Mol. Cell. Proteomics. 11 (2012) M111.010587-M111.010587. doi:10.1074/mcp.M111.010587.

[46] B. Ma, Novor: Real-Time Peptide de Novo Sequencing Software, J. Am. Soc. Mass Spectrom. 26 (2015) 1885–1894. doi:10.1007/s13361-015-1204-0.

[47] U. Keich, W.S. Noble, On the importance of well-calibrated scores for identifying shotgun proteomics spectra, J. Proteome Res. 14 (2015) 1147–1160. doi:10.1021/pr5010983.

[48] A. Keller, A.I. Nesvizhskii, E. Kolker, R. Aebersold, Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search, Anal. Chem. 74 (2002) 5383–5392. doi:10.1021/ac025747h.

[49] A.I. Nesvizhskii, A. Keller, E. Kolker, R. Aebersold, A statistical model for identifying proteins by tandem mass spectrometry, Anal. Chem. 75 (2003) 4646–4658. doi:10.1021/ac0341261.

[50] J.E. Elias, S.P. Gygi, Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry, Nat. Methods. 4 (2007) 207–214.

doi:10.1038/nmeth1019.

[51] R.E. Moore, M.K. Young, T.D. Lee, Qscore: An algorithm for evaluating SEQUEST database search results, J. Am. Soc. Mass Spectrom. 13 (2002) 378–386. doi:10.1016/S1044-0305(02)00352-5.

[52] H. Choi, A.I. Nesvizhskii, Semisupervised model-based validation of peptide identifications in mass spectrometry-based proteomics, J. Proteome Res. 7 (2008) 254–265. doi:10.1021/pr070542g.

[53] L. Käll, J.D. Canterbury, J. Weston, W.S. Noble, M.J. MacCoss, Semi-supervised learning for peptide identification from shotgun proteomics datasets., Nat. Methods. 4 (2007) 923–925. doi:10.1038/nmeth1113.

[54] L. Reiter, M. Claassen, S.P. Schrimpf, M. Jovanovic, A. Schmidt, J.M. Buhmann, M.O. Hengartner, R. Aebersold, Protein Identification False Discovery Rates for Very Large Proteomics Data Sets Generated by Tandem Mass Spectrometry, Mol. Cell. Proteomics. 8 (2009) 2405–2417. doi:10.1074/mcp.M900317-MCP200.

[55] A.K. Shanmugam, A.K. Yocum, A.I. Nesvizhskii, Utility of RNA-seq and GPMDB protein observation frequency for improving the sensitivity of protein identification by tandem MS, J. Proteome Res. 13 (2014) 4113–4119. doi:10.1021/pr500496p.

[56] A.K. Shanmugam, A.I. Nesvizhskii, Effective Leveraging of Targeted Search Spaces for Improving Peptide Identification in Tandem Mass Spectrometry Based Proteomics, J. Proteome Res. 14 (2015) 5169–5178. doi:10.1021/acs.jproteome.5b00504.

[57] K. Ning, A.I. Nesvizhskii, The utility of mass spectrometry-based proteomic data for validation of novel alternative splice forms reconstructed from RNA-Seq data: a preliminary assessment., BMC Bioinformatics. 11 Suppl 1 (2010) S14. doi:10.1186/1471-2105-11-S11-S14.

[58] S. Woo, S.W. Cha, S. Na, C. Guest, T. Liu, R.D. Smith, K.D. Rodland, S. Payne, V. Bafna, Proteogenomic strategies for identification of aberrant cancer peptides using large-scale next-generation sequencing data, Proteomics. 14 (2014) 2719–2730. doi:10.1002/pmic.201400206.

[59] K. Zhang, Y. Fu, W.F. Zeng, K. He, H. Chi, C. Liu, Y.C. Li, Y. Gao, P. Xu, S.M. He, A note on the false discovery rate of novel peptides in proteogenomics, Bioinformatics. 31 (2015) 3249–3253. doi:10.1093/bioinformatics/btv340.

[60] R.M.M. Branca, L.M. Orre, H.J. Johansson, V. Granholm, M. Huss, Å. Pérez-Bercoff, J. Forshed, L. Käll, J. Lehtiö, HiRIEF LC-MS enables deep proteome coverage and unbiased proteogenomics, Nat. Methods. 11 (2014) 59–62. doi:10.1038/nmeth.2732.

[61] P. Abraham, R.M. Adams, G.A. Tuskan, R.L. Hettich, Moving away from the reference genome: Evaluating a peptide sequencing tagging approach for single amino acid polymorphism identifications in the genus populus, J. Proteome Res. 12 (2013) 3642–3651. doi:10.1021/pr400192r.

[62] A.T. Kong, F. V Leprevost, D.M. Avtonomov, D. Mellacheruvu, A.I. Nesvizhskii, MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics, Nat. Publ. Gr. 293 (2017). doi:10.1038/nmeth.4256.

[63] Y. Chen, J. Zhang, G. Xing, Y. Zhao, Mascot-derived false positive peptide identifications revealed by manual analysis of tandem mass spectra, J. Proteome Res. 8 (2009) 3141–3147. doi:10.1021/pr900172v.

[64] S.M. Stevens, K. Prokai-Tatrai, L. Prokai, Factors that contribute to the misidentification of tyrosine nitration by shotgun proteomics., Mol. Cell. Proteomics. 7 (2008) 2442–2451. doi:10.1074/mcp.M800065-MCP200.

[65] J.M. Chick, D. Kolippakkam, D.P. Nusinow, B. Zhai, R. Rad, E.L. Huttlin, S.P. Gygi, A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides, Nat Biotech. 33 (2015) 743–749. doi:10.1038/nbt.3267\rhttp://www.nature.com/nbt/journal/v33/n7/abs/nbt.3267.html#supplementary-information.

[66] R.T. Lawrence, E.M. Perez, D. Hernández, C.P. Miller, K.M. Haas, H.Y. Irie, S.I. Lee, A.C. Blau, J. Villén, The Proteomic Landscape of Triple-Negative Breast Cancer, Cell Rep. 11 (2015) 630–644. doi:10.1016/j.celrep.2015.03.050.

[67] E.W. Deutsch, L. Mendoza, D. Shteynberg, T. Farrah, H. Lam, N. Tasman, Z. Sun, E. Nilsson, B. Pratt, B. Prazen, J.K. Eng, D.B. Martin, A.I. Nesvizhskii, R. Aebersold, A guided tour of the Trans-Proteomic Pipeline, Proteomics. 10 (2010) 1150–1159. doi:10.1002/pmic.200900375.

[68] R. Craig, J.P. Cortens, R.C. Beavis, Open source system for analyzing, validating, and storing protein identification data, J. Proteome Res. 3 (2004) 1234–1242. doi:10.1021/pr049882h.

[69] D. Shteynberg, E.W. Deutsch, H. Lam, J.K. Eng, Z. Sun, N. Tasman, L. Mendoza, R.L. Moritz, R. Aebersold, a. I. Nesvizhskii, iProphet: Multi-level Integrative Analysis of Shotgun Proteomic Data Improves Peptide and Protein Identification Rates and Error Estimates, Mol. Cell. Proteomics. 10 (2011) M111.007690-M111.007690. doi:10.1074/mcp.M111.007690.

[70] D. Shteynberg, A.I. Nesvizhskii, R.L. Moritz, E.W. Deutsch, Combining Results of Multiple Search Engines in Proteomics, Mol. Cell. Proteomics. 12 (2013) 2383–2393. doi:10.1074/mcp.R113.027797.

[71] M. Yadav, S. Jhunjhunwala, Q.T. Phung, P. Lupardus, J. Tanguay, S. Bumbaca, C. Franci, T.K. Cheung, J. Fritsche, T. Weinschenk, Z. Modrusan, I. Mellman, J.R. Lill, L. Delamarre, Predicting immunogenic tumour mutations by combining mass spectrometry and exome sequencing., Nature. 515 (2014) 572–6. doi:10.1038/nature14001.

[72] J. Liepe, F. Marino, J. Sidney, A. Jeko, D.E. Bunting, A. Sette, P.M. Kloetzel, M.P.H. Stumpf, A.J.R. Heck, M. Mishto, A large fraction of HLA class I ligands are proteasome-generated spliced peptides, Science (80-.). 354 (2016) 605–610.

doi:10.1126/science.aaf4384.

[73] M.M. Savitski, ModifiComb, a New Proteomic Tool for Mapping Substoichiometric Post-translational Modifications, Finding Novel Types of Modifications, and Fingerprinting Complex Protein Mixtures, Mol. Cell. Proteomics. 5 (2006) 935–948. doi:10.1074/mcp.T500034-MCP200.

[74] E. Ahrné, F. Nikitin, F. Lisacek, M. Müller, QuickMod: A tool for open modification spectrum library searches, J. Proteome Res. 10 (2011) 2913–2921. doi:10.1021/pr200152g.

[75] C.W.M. Ma, H. Lam, Hunting for unexpected post-translational modifications by spectral library searching with tier-wise scoring, J. Proteome Res. 13 (2014) 2262–2271. doi:10.1021/pr401006g.

[76] A.I. Nesvizhskii, Dynamic Spectrum Quality Assessment and Iterative Computational Analysis of Shotgun Proteomic Data: Toward More Efficient Identification of Post-translational Modifications, Sequence Polymorphisms, and Novel Peptides, Mol. Cell. Proteomics. 5 (2005) 652–670. doi:10.1074/mcp.M500319-MCP200.

[77] B. Bogdanow, H. Zauber, M. Selbach, Systematic Errors in Peptide and Protein Identification and Quantification by Modified Peptides., Mol. Cell. Proteomics. 15 (2016) 2791–801. doi:10.1074/mcp.M115.055103.

[78] D.P. Zolg, M. Wilhelm, K. Schnatbaum, J. Zerweck, T. Knaute, B. Delanghe, D.J. Bailey, S. Gessulat, H.-C. Ehrlich, M. Weininger, P. Yu, J. Schlegl, K. Kramer, T. Schmidt, U. Kusebauch, E.W. Deutsch, R. Aebersold, R.L. Moritz, H. Wenschuh, T. Moehring, S. Aiche, A. Huhmer, U. Reimer, B. Kuster, Building ProteomeTools based on a complete synthetic human proteome, Nat. Methods. 14 (2017). doi:10.1038/nmeth.4153.

[79] Y. Wang, F. Yang, P. Wu, D. Bu, S. Sun, OpenMS-Simulator: an open-source software for theoretical tandem mass spectrum prediction, BMC Bioinformatics. 16 (2015) 110. doi:10.1186/s12859-015-0540-1.

[80] R.D. Bjornson, N.J. Carriero, C. Colangelo, M. Shifman, K.H. Cheung, P.L. Miller, K. Williams, X!!Tandem, an improved method for running X!Tandem in parallel on collections of commodity computers, J. Proteome Res. 7 (2008) 293–299. doi:10.1021/pr0701198.

[81] B. Pratt, J.J. Howbert, N.I. Tasman, E.J. Nilsson, Mr-Tandem: Parallel x!Tandem using Hadoop MapReduce on Amazon web services, Bioinformatics. 28 (2012) 136–137. doi:10.1093/bioinformatics/btr615.

[82] D.T. Duncan, R. Craig, A.J. Link, Parallel tandem: A program for parallel processing of tandem mass spectra using PVM or MPI and X!Tandem, J. Proteome Res. 4 (2005) 1842–1847. doi:10.1021/pr050058i.

[83] J.A. Milloy, B.K. Faherty, S.A. Gerber, Tempest: GPU-CPU computing for high-throughput database spectral matching, J. Proteome Res. 11 (2012) 3581–3591.

doi:10.1021/pr300338p.

[84] L.A. Baumgardner, A.K. Shanmugam, H. Lam, J.K. Eng, D.B. Martin, Fast parallel tandem mass spectral library searching using GPU hardware acceleration, J. Proteome Res. 10 (2011) 2882–2888. doi:10.1021/pr200074h.

[85] Y. Li, H. Chi, L. Xia, X. Chu, Accelerating the scoring module of mass spectrometry-based peptide identification using GPUs., BMC Bioinformatics. 15 (2014) 121. doi:10.1186/1471-2105-15-121.

[86] J.K. Eng, B. Fischer, J. Grossmann, M.J. MacCoss, A fast SEQUEST cross correlation algorithm, J. Proteome Res. 7 (2008) 4598–4602. doi:10.1021/pr800420s.

[87] C.Y. Park, A.A. Klammer, L. Ka, M.J. Maccoss, W.S. Noble, Rapid and Accurate Peptide Identification from Tandem Mass Spectra research articles, J. Proteome Res. 7 (2008) 3022–3027. doi:10.1021/pr800127y.

[88] L.-H. Wang, D.-Q. Li, Y. Fu, H.-P. Wang, J.-F. Zhang, Z.-F. Yuan, R.-X. Sun, R. Zeng, S.-M. He, W. Gao, pFind 2.0: a software package for peptide and protein identification via tandem mass spectrometry., Rapid Commun. Mass Spectrom. 21 (2007) 2985–2991. doi:10.1002/rcm.3173.

[89] B.J. Diament, W.S. Noble, Faster SEQUEST searching for peptide identification from tandem mass spectra, J. Proteome Res. 10 (2011) 3871–3879. doi:10.1021/pr101196n.

[90] N. Satish, M. Harris, M. Garland, {D}esigning {E}fficient {S}oring {A}lgorithms for {M}anycore {GPU}s, Proc. 23rd IEEE Int. Parallel Distrib. Process. Symp. (2009) 1–10.

[91] D.M. Avtonomov, A. Raskind, A.I. Nesvizhskii, BatMass: A Java software platform for LC-MS data visualization in proteomics and metabolomics, J. Proteome Res. 15 (2016) 2500–2509. doi:10.1021/acs.jproteome.6b00021.

[92] D. Fenyö, R.C. Beavis, A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes, Anal. Chem. 75 (2003) 768–774. doi:10.1021/ac0258709.

[93] H. Choi, A.I. Nesvizhskii, False discovery rates and related statistical concepts in mass spectrometry-based proteomics, J. Proteome Res. 7 (2008) 47–50. doi:10.1021/pr700747q.

[94] H. Chi, K. He, B. Yang, Z. Chen, R.X. Sun, S.B. Fan, K. Zhang, C. Liu, Z.F. Yuan, Q.H. Wang, S.Q. Liu, M.Q. Dong, S.M. He, pFind-Alioth: A novel unrestricted database search algorithm to improve the interpretation of high-resolution MS/MS data, J. Proteomics. 129 (2015) 33–41. doi:10.1016/j.jprot.2015.07.019.

[95] S. McIlwain, K. Tamura, A. Kertesz-Farkas, C.E. Grant, B. Diament, B. Frewen, J.J. Howbert, M.R. Hoopmann, L. K??ll, J.K. Eng, M.J. MacCoss, W.S. Noble, Crux: Rapid open source protein tandem mass spectrometry analysis, J. Proteome Res. 13 (2014) 4488–4491. doi:10.1021/pr500741y.

[96] Y. Fu, X. Qian, Transferred Subgroup False Discovery Rate for Rare Post-translational

Modifications Detected by Mass Spectrometry., Mol. Cell. Proteomics. 13 (2014) 1359–68. doi:10.1074/mcp.O113.030189.

[97] M. Vaudel, J.M. Burkhart, R.P. Zahedi, E. Oveland, F.S. Berven, A. Sickmann, L. Martens, H. Barsnes, PeptideShaker enables reanalysis of MS-derived proteomics data sets, Nat. Biotechnol. 33 (2015) 22–24. doi:10.1038/nbt.3109.

[98] J. V Olsen, M. Mann, Status of large-scale analysis of post-translational modifications by mass spectrometry., Mol. Cell. Proteomics. 12 (2013) 3444–52. doi:10.1074/mcp.O113.034181.

[99] K. Sharma, R.C.J. D'Souza, S. Tyanova, C. Schaab, J. Wiśniewski, J. Cox, M. Mann, Ultradeep Human Phosphoproteome Reveals a Distinct Regulatory Nature of Tyr and Ser/Thr-Based Signaling, Cell Rep. 8 (2014) 1583–1594. doi:10.1016/j.celrep.2014.07.036.

[100] Y. Perez-Riverol, E. Alpi, R. Wang, H. Hermjakob, J.A. Vizca??no, Making proteomics data accessible and reusable: Current state of proteomics databases and repositories, Proteomics. 15 (2015) 930–950. doi:10.1002/pmic.201400302.

[101] O.S. Skinner, N.L. Kelleher, Illuminating the dark matter of shotgun proteomics, Nat. Biotechnol. 33 (2015) 717–718. doi:10.1038/nbt.3287.

[102] Y. Pozniak, N. Balint-Lahat, J.D. Rudolph, C. Lindskog, R. Katzir, C. Avivi, F. Pontén, E. Ruppin, I. Barshack, T. Geiger, System-wide Clinical Proteomics of Breast Cancer Reveals Global Remodeling of Tissue Homeostasis, Cell Syst. 2 (2016) 172–184. doi:10.1016/j.cels.2016.02.001.

[103] E.L. Huttlin, L. Ting, R.J. Bruckner, F. Gebreab, M.P. Gygi, J. Szpyt, S. Tam, G. Zarraga, G. Colby, K. Baltier, R. Dong, V. Guarani, L.P. Vaites, A. Ordureau, R. Rad, B.K. Erickson, M. Wühr, J. Chick, B. Zhai, D. Kolippakkam, J. Mintseris, R.A. Obar, T. Harris, S. Artavanis-Tsakonas, M.E. Sowa, P. De Camilli, J.A. Paulo, J.W. Harper, S.P. Gygi, The BioPlex Network: A Systematic Exploration of the Human Interactome, Cell. 162 (2015) 425–440. doi:10.1016/j.cell.2015.06.043.

[104] K. Kramer, T. Sachsenberg, B.M. Beckmann, S. Qamar, K.-L. Boon, M.W. Hentze, O. Kohlbacher, H. Urlaub, Photo-cross-linking and high-resolution mass spectrometry for assignment of RNA-binding sites in RNA-binding proteins, Nat. Methods. 11 (2014) 1064–1070. doi:10.1038/nmeth.3092.

[105] C.-C. Tsou, D. Avtonomov, B. Larsen, M. Tucholska, H. Choi, A.-C. Gingras, A.I. Nesvizhskii, DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics, Nat. Methods. 12 (2015) 258–264. doi:10.1038/nmeth.3255.

[106] F. Kryuchkov, T. Verano-Braga, T.A. Hansen, R.R. Sprenger, F. Kjeldsen, Deconvolution of mixture spectra and increased throughput of peptide identification by utilization of intensified complementary ions formed in tandem mass spectrometry, J. Proteome Res. 12 (2013) 3362–3371. doi:10.1021/pr400210m.

[107] S. Houel, R. Abernathy, K. Renganathan, K. Meyer-Arendt, N.G. Ahn, W.M. Old, Quantifying the impact of chimera MS/MS spectra on peptide identification in large-scale proteomics studies, J. Proteome Res. 9 (2010) 4152–4160. doi:10.1021/pr1003856.

[108] B. Zhang, M. Pirmoradian, A. Chernobrovkin, R. a Zubarev, DeMix Workflow for Efficient Identification of Co-fragmented Peptides in High Resolution Data-dependent Tandem Mass Spectrometry., Mol. Cell. Proteomics. 13 (2014) 3211–3223. doi:10.1074/mcp.O114.038877.

[109] B. Metz, Identification of Formaldehyde-induced Modifications in Proteins: REACTIONS WITH MODEL PEPTIDES, J. Biol. Chem. 279 (2003) 6235–6243. doi:10.1074/jbc.M310752200.

[110] O. Kabil, R. Banerjee, Enzymology of H2S biogenesis, decay and signaling., Antioxid. Redox Signal. 20 (2014) 770–82. doi:10.1089/ars.2013.5339.

[111] D.W. Huang, R. a Lempicki, B.T. Sherman, Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources., Nat. Protoc. 4 (2009) 44–57. doi:10.1038/nprot.2008.211.

[112] H. Choi, B. Larsen, Z.-Y. Lin, A. Breitkreutz, D. Mellacheruvu, D. Fermin, Z.S. Qin, M. Tyers, A.-C. Gingras, A.I. Nesvizhskii, SAINT: probabilistic scoring of affinity purification-mass spectrometry data., Nat. Methods. 8 (2011) 70–3. doi:10.1038/nmeth.1541.

[113] M.E. Sardiu, M.P. Washburn, Construction of protein interaction networks based on the label-free quantitative proteomics, Methods Mol Biol. 781 (2011) 71–85. doi:10.1007/978-1-61779-276-2_5.

[114] I. Van Den Broek, F.P.H.T.M. Romijn, N.P.M. Smit, A. Van Der Laarse, J.W. Drijfhout, Y.E.M. Van Der Burgt, C.M. Cobbaert, Quantifying protein measurands by peptide measurements: Where do errors arise?, J. Proteome Res. 14 (2015) 928–942. doi:10.1021/pr5011179.

[115] M. Tan, H. Luo, S. Lee, F. Jin, J.S. Yang, E. Montellier, T. Buchou, Z. Cheng, S. Rousseaux, N. Rajagopal, Z. Lu, Z. Ye, Q. Zhu, J. Wysocka, Y. Ye, S. Khochbin, B. Ren, Y. Zhao, Identification of 67 histone marks and histone lysine crotonylation as a new type of histone modification, Cell. 146 (2011) 1016–1028. doi:10.1016/j.cell.2011.08.008.

[116] P. Mertins, D.R. Mani, K. V. Ruggles, M.A. Gillette, K.R. Clauser, P. Wang, X. Wang, J.W. Qiao, S. Cao, F. Petralia, E. Kawaler, F. Mundt, K. Krug, Z. Tu, J.T. Lei, M.L. Gatza, M. Wilkerson, C.M. Perou, V. Yellapantula, K. Huang, C. Lin, M.D. McLellan, P. Yan, S.R. Davies, R.R. Townsend, S.J. Skates, J. Wang, B. Zhang, C.R. Kinsinger, M. Mesri, H. Rodriguez, L. Ding, A.G. Paulovich, D. Fenyö, M.J. Ellis, S.A. Carr, C. Nci, Proteogenomics connects somatic mutations to signalling in breast cancer, Nature. 534 (2016) 55–62. doi:10.1038/nature18003.

[117] H. Zhang, T. Liu, Z. Zhang, S.H. Payne, B. Zhang, J.E. McDermott, J.Y. Zhou, V.A. Petyuk, L. Chen, D. Ray, S. Sun, F. Yang, L. Chen, J. Wang, P. Shah, S.W. Cha, P.

Aiyetan, S. Woo, Y. Tian, M.A. Gritsenko, T.R. Clauss, C. Choi, M.E. Monroe, S. Thomas, S. Nie, C. Wu, R.J. Moore, K.H. Yu, D.L. Tabb, D. Feny??, V. Vineet, Y. Wang, H. Rodriguez, E.S. Boja, T. Hiltke, R.C. Rivers, L. Sokoll, H. Zhu, I.M. Shih, L. Cope, A. Pandey, B. Zhang, M.P. Snyder, D.A. Levine, R.D. Smith, D.W. Chan, K.D. Rodland, S.A. Carr, M.A. Gillette, K.R. Klauser, E. Kuhn, D.R. Mani, P. Mertins, K.A. Ketchum, R. Thangudu, S. Cai, M. Oberti, A.G. Paulovich, J.R. Whiteaker, N.J. Edwards, P.B. McGarvey, S. Madhavan, P. Wang, D.W. Chan, A. Pandey, I.M. Shih, H. Zhang, Z. Zhang, H. Zhu, L. Cope, G.A. Whiteley, S.J. Skates, F.M. White, D.A. Levine, E.S. Boja, C.R. Kinsinger, T. Hiltke, M. Mesri, R.C. Rivers, H. Rodriguez, K.M. Shaw, S.E. Stein, D. Fenyo, T. Liu, J.E. McDermott, S.H. Payne, K.D. Rodland, R.D. Smith, P. Rudnick, M. Snyder, Y. Zhao, X. Chen, D.F. Ransohoff, A.N. Hoofnagle, D.C. Liebler, M.E. Sanders, Z. Shi, R.J.C. Slebos, D.L. Tabb, B. Zhang, L.J. Zimmerman, Y. Wang, S.R. Davies, L. Ding, M.J.C. Ellis, R.R. Townsend, Integrated Proteogenomic Characterization of Human High-Grade Serous Ovarian Cancer, Cell. 166 (2016) 755–765. doi:10.1016/j.cell.2016.05.069.

[118] G.P. Mommen, C.K. Frese, H.D. Meiring, J. van Gaans-van den Brink, A.P. de Jong, C.A. van Els, A.J. Heck, Expanding the detectable HLA peptide repertoire using electron-transfer/higher-energy collision dissociation (EThcD), Proc Natl Acad Sci U S A. 111 (2014) 4507–4512. doi:10.1073/pnas.1321458111.