Published in final edited form as:
*Anal Chem*. 2010 August 15; 82(16): 6821–6829. doi:10.1021/ac100783x.

# MacroSEQUEST: Efficient candidate-centric searching and high resolution correlation analysis for large-scale proteomics datasets

**Brendan K. Faherty**[1] and **Scott A. Gerber**[1,2,*]

[1] Department of Genetics, Dartmouth Medical School, Lebanon, NH 03756

[2] Norris Cotton Cancer Center, Lebanon, NH 03756

## Abstract

Modern mass spectrometers are now capable of producing tens of thousands of tandem mass (MS/MS) spectra per hour of operation, resulting in an ever-increasing burden on the computational tools required to translate these raw MS/MS spectra into peptide sequences. In the present work, we describe our efforts to improve the performance of one of the earliest and most commonly used algorithms, SEQUEST, through a wholesale redesign of its processing architecture. We call this new program MacroSEQUEST, which exhibits a dramatic improvement in processing speed by transiently indexing the array of MS/MS spectra prior to searching FASTA databases. We demonstrate the performance of MacroSEQUEST relative to a suite of other programs commonly encountered in proteomics research. We also extend the capability of SEQUEST by implementing a parameter in MacroSEQUEST that allows for scalable sparse arrays of experimental and theoretical spectra to be implemented for high resolution correlation analysis, and demonstrate the advantages of high-resolution MS/MS searching to the sensitivity of large-scale proteomics datasets.

## 1. Introduction

Mass spectrometry (MS) coupled with computer-assisted database spectral matching has evolved into a cornerstone technology that drives research for the field of proteomics1. Recent developments in MS instrumentation have resulted in commercial mass spectrometers capable of generating tens of thousands of tandem mass spectra (MS/MS) per penultimate online reverse-phase liquid chromatography (LC) separation2¯4. When coupled with biochemical prefractionation methods, a complete dataset for a proteomics experiment (including technical and biological replicates) can consist of millions of MS/MS spectra per biological sample. Importantly, although the absolute sensitivity of new instruments to detect a single peptide species has improved, it has been suggested that much of the credit for increased depth of peptide and protein coverage, and improved detection of sub-stoichiometric species belongs to the increased rate of MS/MS spectral acquisition of these instruments, which allows them to penetrate transient rasters of precursor ions to greater ion peak depth per chromatographic unit time5. Indeed, the number of candidate precursor ions (MS1 features) is often at least an order of magnitude greater than the number of MS/MS sequencing events in a typical LC-MS analysis6. Given the current level of success with this strategy, it is only reasonable to predict that this trend of increasing MS/MS bandwidth to improve overall peptide identification rates per sample will continue, which places an additional burden on the computational tools required to search these larger datasets.

[*] to whom correspondence should be addressed: scott.a.gerber@dartmouth.edu.

A single raw MS/MS spectrum is translated into a peptide spectral match (PSM) through the use of algorithms that first search translated genomic databases from an organism of interest for candidate peptides by a defined enzyme specificity (based on the protease used for digestion) and precursor mass (based on the MS1 feature mass from which the MS/MS spectrum was derived, and a desired mass precision)[7]. Other parameters (fixed protein modifications, variable post-translational modifications, number of missed enzyme cleavage loci, maximum and/or minimum peptide size, etc.) may also be included. This search results in a list of candidate peptides that may contain the correct peptide sequence from which the MS/MS spectrum was derived. These candidate peptides are then evaluated for correctness by comparing the observed MS/MS spectrum with a dynamically generated theoretical spectrum for each candidate peptide and given a score that reflects the quality of their match. These scores are then ranked, and in some cases, further evaluated[8], before reporting the “best” candidate peptide match (PSM) for the MS/MS spectrum. In a collection of MS/MS spectra from a single LC-MS run, a series of LC-MS runs for a given sample, or from a series of different samples, this general process is iterated thousands, if not millions, of times.

Today, the modern proteomics researcher can choose from several computational tools to translate MS/MS spectra into PSMs, including SEQUEST[9], Mascot[10], X!Tandem[11, 12], and OMSSA[13], among others. A number of studies have been performed that evaluate the analytical performance of these algorithms, with regards to precision and sensitivity of the PSM collections they generate[14–16]. While it appears that certain algorithms perform slightly better or worse than others based on the nature of the sample (e.g. phosphorylation, enzyme digest), the type of mass spectrometer used to generate these MS/MS data, and the mechanism of peptide fragmentation, etc., they are in general more similar than they are different – at least, in terms of the nature of the PSM collections they produce. However, given the recent trend towards larger and larger numbers of MS/MS spectra per experiment, the relative processing speed of these algorithms has become an important practical consideration, as some of these algorithms perform significantly slower than others. In particular, SEQUEST has lagged significantly behind, although limited efforts to improve performance by parallelization have been reported[17]. In general, the most basic solution to any large discrepancies in performance (or “productivity”, defined as the number of MS/MS spectra processed per unit computational time) has been to index the protein database by pre-digesting the protein sequences to peptides, then indexing each candidate peptide by precursor mass. While these peptide indices do result in significant performance gains that in general level the playing field among algorithms, there are drawbacks to using indexed databases. For example, the use of an indexed database as opposed to the use of a FASTA database results in a significant expansion of disk memory, requires separate indices for each enzyme used for digestion, and is problematic when considering post-translational modifications, all of which is repeated and exacerbated for each organism and/or release version of the organism-specific genome sequence. A recent report attempted to reconcile the issues of both performance and indexed database files by reading the target FASTA database once for the first MS/MS spectrum to be analyzed, indexing “on-the-fly” and storing this peptide index in fast CPU memory for use in finding candidates for subsequent MS/MS spectra[18].

An alternative approach that addresses the performance and convenience issues associated with both indexed database files as well as indexing the protein database in CPU memory is to consider an entire set of MS/MS spectra to be searched (e.g. a complete LC-MS/MS run) as an array of precursor masses and spectra, and to transiently index these MS/MS spectra by precursor mass at the beginning of a search. This allows for efficient matching of candidate peptides during a single *in silico* digestion pass through a FASTA database. Because the number of MS/MS spectra in an LC-MS/MS run is small relative to the number of candidate peptides in a typical organism database, this indexing step is extremely fast and can easily be implemented at the launch of a search. Furthermore, additional performance gains can be realized through elimination of redundant candidate peptide spectral processing steps that occur when single MS/MS spectra are considered in the absence of other MS/MS spectra that have overlapping candidate peptide search spaces. Indeed, the benefits of this “candidate-centric” (as opposed to “spectrum-centric”) searching approach have been described explicitly, first in principle by Edwards & Lippert19 and later in practice by Tabb and coworkers20.

In the present work, we demonstrate the general utility of this “candidate-centric” approach by modifying the commercial version of SEQUEST to perform searches in a candidate-centric fashion, resulting in a program we call MacroSEQUEST. MacroSEQUEST reads FASTA-formatted protein databases and returns PSMs for a collection of MS/MS spectra in a fraction of the time it takes legacy SEQUEST. We compare the performance of MacroSEQUEST to a suite of other commonly used database search engines. Finally, we demonstrate a useful application for the performance gains associated with MacroSEQUEST by leveraging this increase in speed to perform high resolution MS/MS correlation analysis, and describe the benefits to classical SEQUEST scores associated with high mass precision fragment ion searching.

## 2. Materials and Methods

### 2.1 Sample preparation

Peptide synthesis was performed by New England Peptide (Gardner, MA). Yeast cells were harvested during logarithmic growth, pelleted and lysed by addition of SDS-PAGE pH 8.1 sample buffer (3x volume buffer:pellet weight) and bead beating at 4 °C. HeLa cell lysate was prepared by harvesting a confluent 15-cm dish of HeLa cells by trypsinization, washing 2x in PBS, and addition of 4 ml lysis buffer (0.5% Triton X-100, 50mM Tris pH 8.1, 150mM NaCl, 1mM $MgCl_2$, Roche Mini-complete protease inhibitors), followed by sonication and clarification of the lysate in a centrifuge (14,000 × *g*) for 10 minutes at 4 °C. Both protein preps were reduced by addition of DTT to 5 mM and incubation in a water bath at 55 °C for 20 minutes, followed by alkylation of cysteines in 12.5 mM iodoacetamide at room temperature for 45 minutes. After quenching the alkylation reaction (addition of 2.5mM DTT), the lysates were aliquotted, snap frozen in liquid nitrogen and stored at −80 °C until use. To prepare the protein digests, an aliquot of cell lysate was warmed rapidly under warm water and mixed with SDS-PAGE sample buffer to a protein concentration of ~0.25mg/ml protein, followed by separation on 2-well, 4 – 12 % NOVEX minigels (for 2D separations) with a protein marker in the narrow lane. Proteins were visualized with Coomassie blue and the region between 80 and 125kDa was excised, destained, and digested with trypsin. Peptide samples were analyzed on an LTQ Orbitrap (ThermoFisher Scientific, Bremen, Germany) per established procedures5. LTQ-Orbitrap .RAW files were converted to .mzXML files using ReAdW.exe (version 4.0, http://sourceforge.net/projects/sashimi/files/). Peptide precursor mass assignments were adjusted post-acquisition with in-house software that i) updates MS2 scan headers with high mass accuracy precursor information from MS1 scans, and ii) averages multiple observations of these precursor values across each peptide chromatographic elution profile.

### 2.2 MacroSEQUEST

MacroSEQUEST was written from scratch in ANSI C using standard libraries and compiled with GCC version 4.1.2 on a Unix platform, using the SEQUEST2.8 source code as a guide. Input is provided as command line arguments. Experimental data is read in .DTA file format, and candidate peptides are read from FASTA-formatted protein databases. Macro outputs SEQUEST-like .out files for each input spectrum. . To measure the efficiency of the “candidate-centric” method, we created a version of Macro (Macro Null) in which the *Xcorr* scoring function was replaced with a random number generator.

### 2.3 Database searching

Searches were performed using the latest database builds from the yeast proteome (Saccharomyces Genome Database, http://www.yeastgenome.org/) or the human proteome (UniProtKB, http://www.uniprot.org/). Target-decoy databases were generated using in-house scripts that reverse each protein sequence, label decoys proteins with a specific identifier and append the decoy sequences to the end of a forward database21. Unless otherwise stated, all searches were performed with a 1.1 Da precursor mass tolerance and filtered to +/− 1.5 ppm precursor mass measurement accuracy (Figure S-1); *Xcorr* and *dCn* cutoff values were adjusted to achieve a false discovery rate (FDR) of less than 1% (Table S-1). Only the yeast searches were conducted with semi-tryptic enzyme specificity; all other searches were performed with full trypsin specificity. The maximum number of missed cleavages for all searches was set to three. All searches were also performed with acetamide-modified Cys as a static modification (+ 57.021461 Da) and with oxidized Met as a variable modification (+/− 15.991915 Da); for phosphorylation searches, Ser, Thr, and Tyr were allowed to vary by +/− 79.966331 Da. Variable modifications were limited to a maximum of 3 per peptide, where applicable. Refinement or iterative searches were not permitted when using X!Tandem and OMSSA. No multithreading was used for any search algorithm.

## 3. Results

### 3.1 Candidate-centric database spectral matching

Historically, database spectral matching algorithms such as SEQUEST dealt with very small numbers of MS/MS spectra acquired per experiment, from as few as tens to at most one hundred spectra in a typical analysis9. Individual MS/MS spectra were then matched with candidate peptides by searching through a target database. In general, the focus for algorithm development has been on the steps involved in this single iteration, and on improving the sensitivity, accuracy, and productivity of a single analysis, with the expectation that this “optimized” core process is iterated until PSMs have been generated for all MS/MS spectra in an analysis which, while computationally inefficient, was adequate given the historical context of the problem space. A generalized scheme describing this classical workflow for SEQUEST is depicted in Figure 1A. However, this workflow fails to recognize potential elements of relatedness between MS/MS spectra that are collected together. Clearly, all of these spectra require matching to candidate peptides in the same database; thus, presenting the entire collection of spectra to the database (or vice-versa) as a “single analysis” has the potential to reduce areas of overhead associated with repeating a single process, such as digesting a database, thousands of times (Figure 1B). Our analysis of the work distribution for SEQUEST when searching a target-decoy human UniProtKB protein database (~150,000 proteins) against 10,000 spectra reveals that only a very small portion of the actual search time is spent on scoring candidate PSMs (~1%), while the remainder is spent parsing the database (Figure 1C). Although the database must be read once in order to generate a PSM for a single spectrum, it does not need to be re-read to search other MS/MS spectra under the same set of parameters if those additional spectra are considered simultaneously. In this way, legacy SEQUEST wastes a significant amount of search time when large numbers of MS/MS spectra are searched against large databases.

We sought to address these issues through a wholesale re-architecting of the SEQUEST scoring components in order to conduct them in a candidate-centric fashion, and named the resulting program MacroSEQUEST, or “Macro” for short. The Macro workflow is described in Figure 1B. Similar to the candidate-centric algorithm DBDigger20, a single Macro process at launch is fed a parameter set that includes a path to a collection of MS/MS spectra. This collection of spectra is first read into memory and indexed by precursor ion mass to create a “spectral data array” that includes preprocessing for each experimental spectrum and memory mapping for candidate peptide sequences, protein references, search scores, etc. After this data structure is created, Macro begins parsing the FASTA database of interest by digesting it based on a desired enzyme cleavage specificity. For each newly digested peptide, Macro calculates its mass and checks it against the spectral array, including a given mass tolerance; if a peptide falls outside these boundaries, Macro discards it and proceeds to the next logical peptide in the database. If, however, a peptide falls into one or more spectral “bins” in the spectral array, Macro enters a scoring loop in which a re-implementation of a fast SEQUEST cross correlation algorithm22, 23 is performed on the candidate peptide for each MS/MS spectrum that falls within the peptide’s desired precursor mass tolerance, and plugs these scores and associated peptide/protein information into the spectral array. The scoring loop then closes by returning to database parsing, and to the next logical peptide in the database; Macro continues this cycle of peptide generation, spectral array scanning and scoring until it reaches the end of the FASTA database. Macro then concludes by calculating final score differences, etc. for the top-ranked candidate PSMs and writing this information to result files. Because Macro was developed from the original SEQUEST source code, the primary scoring metric *Xcorr* in Macro is identical to those created by legacy SEQUEST – a major difference between the two is that Macro no longer performs Sp scoring as a preliminary scoring step but instead calculates *Xcorr* for all candidate PSMs, owing to the computationally fast, non-FFT correlation algorithm. Macro spends significantly less time parsing the target-decoy human database than SEQUEST during a search of the same 10,000 spectra (Figure 1C, inset), and almost 87% of the total run time performing scoring functions.

Because such a small fraction of the actual SEQUEST process is spent scoring candidate PSMs, and because this cycle is repeated in its entirety for each spectrum to be searched, SEQUEST’s productivity (number of spectra searched/unit time spent searching) is relatively flat as a function of the number of spectra searched (Figure 2). Macro, however, benefits from having more spectra to search by distributing the fixed time cost associated with a single digestion and parsing pass through the database across many spectra, until the time spent parsing the database is small relative to the time spent scoring candidate PSMs, at which point Macro’s search productivity flattens out. This improvement in productivity as a function of number of spectra searched is clearly depicted in Figure 2, where searches of a target-decoy yeast (Figure 2A), human (Figure 2B), and human databases with dynamic phosphorylation on serine, threonine, and tyrosine (Figure 2C) using Macro all outperform legacy SEQUEST by 102x, 107x, and 42x, respectively, when searching 10,000 MS/MS spectra. Note that the increase in parsing logic associated with determining dynamic protein post-translational modifications such as phosphorylation substantially compresses and flattens the Macro productivity curve, and extends the point in numbers of spectra at which Macro achieves maximum productivity to well beyond the 10,000 mark.

In order to assess the level of spectral array processing overhead in Macro, we generated a “Macro Null” version of the program that replaces the *XCorr* scoring functions with a simple random number generator. This allowed us to subtract the time it takes Macro to parse the database and scan the array of spectra from the total search time, which generated an ideal power fit as a function of the number of spectra searched (Figure 2A – C).

### 3.2 Comparison of MacroSEQUEST to other contemporary tools

Although it was among the earliest algorithms to be adopted into widespread use, SEQUEST is now not the only program available to retrieve candidate PSMs from FASTA databases. Other commercial (Mascot) and non-commercial (OMSSA, X!Tandem) programs are also designed to execute database searches, although the core components and matching algorithms differ substantially. To establish Macro's performance rank among programs that are commonly encountered in proteomics research labs and protein identification core facilities, we generated productivity curves for searches against our target-decoy human database for Macro, SEQUEST, Mascot, OMSSA, and X!Tandem. We also noted that all three of these programs have been written to take advantage of multiple physical and logical cores now increasingly common in modern CPU architectures, including the Intel Conroe chip used in this comparison. In order to cleanly reconcile differences in performance due to multithreading, we disabled one of the two cores on our Conroe prior to performing the comparison, the results of which are depicted in Figure 3. At peak productivity, Macro performs equivalent to OMSSA, about 20% faster than Mascot, and twice as fast X!Tandem, with legacy SEQUEST lagging far behind the others.

### 3.3 High resolution MS/MS correlation analysis

In light of the significant performance improvement of Macro over legacy SEQUEST, we reasoned that this additional speed could be leveraged to also produce gains in the quality of spectral matches by creating sparse arrays of MS/MS fragment ions from high resolution, high mass accuracy instruments such as the LTQ Orbitrap2 and, in particular, the LTQ Orbitrap Velos3, and by performing full correlation analyses on them to produce *Xcorr* scores, a feature not currently available in SEQUEST. Given the novelty of *Xcorr* and *dCn* values derived from high resolution MS/MS spectra, we were also interested in evaluating the qualitative impact that variable search "resolution" might have on these scores.

To do this, we created a parameter for Macro that defines the bin width of the arrays (in *m/z*) used for correlation analysis, and used this factor to scale the number of bins that are generated within the range of observed fragment ions during MS/MS spectrum pre-processing steps. Macro then populates these bins with normalized ion intensity values using logic consistent with legacy SEQUEST, and performs the fast *Xcorr* pre-processing math on the resultant sparse array. Given the mass precision and resolution of ion traps, MS/MS data from such instruments normally defines these bin widths as 1 m/z wide, and the spectral preprocessing logic of SEQUEST determines which ions fall into which *m/z* bin(s). Figure 4A depicts an MS/MS spectrum collected in an LTQ Orbitrap with a resolution setting of 15,000. Figure 4B describes this spectrum pre-processed for fast *Xcorr* analysis by Macro with a bin width of 1 *m/z*, while Figure 4C depicts an array from the same spectrum but with bin widths of 0.025 *m/z*. Note that, in the insets, binning logic in Macro reduced the number of available ions for matching across a 2 *m/z*-wide window in the original spectrum from six to two when 1 *m/z*-wide bins were used versus 0.01 *m/z*-wide bins. Clearly, the higher "resolution" of the array in Figure 4C allows for more accurate ion-to-bin assignments, and consequently a more robust discrimination between ion fragments of similar *m/z*. This is also apparent in the significant reduction in the average magnitude of negative values in the high resolution pre-processed array (Figure 4C), which spreads these anti-correlations out over a much larger bin space. Similar to the pre-processing steps that were performed for experimental spectra, theoretical spectra for each candidate peptide were created using the same bin width scaling factor, and *Xcorr* values were then calculated from the dot product of the two arrays.

We then tested the utility of these scalable MS/MS arrays by collecting data from our LTQ Orbitrap at relatively high resolution and mass accuracy (R = 15,000). We analyzed our HeLa cell 80 – 120kDa protein lysate digest over a 90-minute LTQ Orbitrap gradient, during which a total of 5,725 MS/MS spectra were collected. We then used Macro to iteratively search this run with fragment ion bin widths of 1.0, 0.75, 0.5, 0.25, 0.1, 0.075, 0.05, 0.025 and 0.01 *m/*z. Because the size of the arrays scales inversely with bin width, there is a modest but significant speed penalty associated with high resolution scoring. On our Intel Conroe test platform fitted with 8Gb RAM, we observed a 3x increase in search time using bin widths of 0.025 *m/z* relative to unit resolution, and a 6x increase in search time when using bin widths of 0.01 *m/z*. However, we also observed a significant increase in the total number of true positive (TP) PSMs at a 1% false discovery rate (FDR). Figure 5A depicts the total number of TP PSMs for this run as a function of search bin width. We observed a 24% increase in the number of TP PSMs when using bin widths of 0.025 *m/z* versus 1.0 *m/z* bin widths. We considered the possibility that simply limiting the number of candidate PSMs by using a very narrow precursor ion tolerance may allow these “rescuable” PSMs to rise in rank. However, consistent with previous reports[24, 25], we only observed a very slight increase in TP assignments in a search limited to +/− 10ppm precursor ion mass measurement accuracy (Figure 5B). This suggests that the sensitivity gains we observe by high resolution fragment ion searching represent PSMs that are not accessible by simply limiting the number of candidate PSMs in a search – that although these mediocre yet ”rescuable” MS/MS spectra are being considered in the context of fewer competing candidates during a narrow window precursor ion search, their inherently poor correlation characteristics do not sufficiently enhance their differential *Xcorr* rank to allow for their rescue in this search mode.

This increase in true positive assignments can be ascribed to a combination of features of merit associated with the higher resolution versions of *Xcorr* and *dCn* (delta-correlation). Although we did not see an appreciable change in *Xcorr* for TP PSMs when searching at higher fragment ion resolution, we noted that false positive (FP) PSMs exhibited a decrease in their median *Xcorr* values (from 1.6 to 1.2 for 1.0 and 0.025 *m/z*-wide bins, respectively; Figure 5C). This allows for significantly lower *Xcorr* cutoff values to be used when establishing a 1% FDR. From Figure 5C, clearly many of the “rescued” TP PSMs come from low *Xcorr* scores that are otherwise inaccessible due to the extension of the FP PSM distribution into that score space. Equally dramatic was the change in TP PSM *dCn* (delta-correlation) scores, whose median values increased from 0.25 to 0.42 across the entire dataset (Figure 5D); for those “rescued” TP PSMs, the median *dCn* value jumped from 0.06 to 0.33. As *dCn* is a direct measure of the relative *Xcorr* score separation between the highest-ranked PSM and its nearest neighbor (*de facto* a FP PSM), our observations of overall reduced FP *Xcorr* scores and increased TP *dCn* scores are consistent with this relationship.

To further describe the nature of some of these “rescued” TP PSMs, we examined their scores and individual PSM rankings. Figure 6A depicts the top two ranked PSMs for a particular MS/MS spectrum that yielded an excellent *Xcorr* score (4.3), but a very low *dCn* score of 0.005 when scored with a 1.0 *m/z* bin width. Note that these PSMs differ only by a K/Q (mass difference = 0.036 Da) in the middle of the peptide sequences. Although distinguishable by Orbitrap precursor ion mass precision, for correlation analysis at unit resolution fragment ion correlation, this difference is transparent and results in almost identical *Xcorr* scores for these two sequences. However, when searched by Macro using 0.025 *m/z* bin widths, this K/Q difference is readily distinguished by *dCn*, likely for all singly and doubly-charged fragment ions from this quadruply-charged precursor that contain the Lys residue. Figure 6B describes the general behavior of all candidate PSMs for this MS/MS spectrum at the two fragment ion tolerances. It is worth noting again in this plot that the discriminating power of these high resolution correlations lies predominantly in positive overlap between the theoretical and experimental arrays, and not in the “signal-to-noise” of this overlap relative to neighboring offsets (+/− 75 *m/z* equivalents) – an important aspect of the unit resolution *Xcorr*. This can be observed in the large decrease in the number of negative *Xcorr* values for low-ranked PSMs between the high and low resolution searches.

Indeed, of the 819 “rescued” PSMs, 26% of them were for peptide sequences that were correctly assigned (top-ranked) in the lower resolution correlation analysis, but with extremely low *dCn*/*Xcorr* combinations such that they precluded a definitive assignment when score cutoffs were applied to achieve a 1% FDR. The remaining rescued peptide matches were re-ranked from lower ranks to the top-ranked candidate PSM. Figure 6C depicts the distribution across the entire run of original candidate PSM ranks when searches were performed with 1.0 *m/z*-wide fragment ion bins that are “rescued” when the search is done at 0.025 *m/z* bin widths. Of those PSMs that are re-ranked between the two searches, the median rank in the low-resolution search is 9 – the lowest re-ranked candidate PSM moved from the 4,517th position to the top-ranked spot when searched at high resolution. Although this MS/MS spectrum is assigned a PSM (AASVHTVGEDTEETPHR; $M+4H^+$, MMA = 0.1 ppm) with an *Xcorr* of only 0.67, the *dCn* is 0.08 (*dCn* = 0.001 @ unit fragment ion resolution). This peptide is from the protein hPOP1 with a molecular weight of 115kDa, which is in the range of molecular weights (80 – 125kDa) of SDS-PAGE-fractionated human cell lysate from which the sample was derived. Figure 6D displays a reciprocal plot of the endogenous, rescued MS/MS spectrum from lysate with the matched ions noted and a mirrored spectrum from a synthetic peptide that we analyzed under identical LC-MS/MS conditions, confirming this rescued sequence as being a correct match. Additionally, there were two other peptides from the same protein that were identified in both the high and low resolution correlation analyses with scores and mass measurement accuracy assignments that pass cutoffs (KTHQPSDEVGTSIEHPR, $M+4H^+$, *Xcorr* = 3.78, *dCn* = 0.43, MMA = −0.2 ppm and IPILLIQQPGK, $M+2H^+$, *Xcorr* = 3.31, *dCn* = 0.51, MMA = 0.1 ppm), supporting the likelihood that this PSM is a valid match.

## 4. Discussion

Soon, proteomics researchers will be swimming in a sea of mass spectra. During the later stages of development of Macro, we were asked to search a 5-hour LTQ Velos run with Macro, and were surprised to find that it consisted of ~170,000 tandem mass spectra. Clearly, the relative processing speed of computer-assisted spectral matching algorithms will be a critical feature for these algorithms in the coming years. Our intent with developing Macro was to refresh a scoring metric (*Xcorr*) that adds significant value to proteomics research by reorganizing the way it conducted searches. That Macro is also very competitive with other search engines reinforces our view that candidate-centric searching represents a practical approach to search algorithm design in general. The core Macro process is also scalable: the MS/MS spectral array may be accessed in main CPU memory by multiple independent processing cores, each of which could digest a different portion of the FASTA database or perform scoring functions, or both – design modifications such as this are planned for future versions of Macro.

We feel that the performance gains associated with Macro can allow for a host of features and functions to be added to the classical scoring function *Xcorr*, including post-translational modification scanning and, as we show here, high resolution correlation analysis. Although a parameter does exist in legacy SEQUEST that appears to allow users to scale fragment ion tolerances, closer inspection of the SEQUEST source code and search output when using this parameter reveals that it is not performing these searches as the user likely intends. Our inclusion of a variable fragment ion bin width parameter that accurately bins fragment ions into a scalable sparse array enables users with a mass spectrometer capable of better than unit resolution MS/MS spectra to search their data with the corresponding fragment ion mass tolerance. This has the added benefit of improved overall search accuracy relative to unit resolution searching. Although this does come as a speed penalty, we note that even when running at 0.025 *m/*z bin width, Macro is still 20x faster than SEQUEST on our platform. We anticipate that these high resolution correlation searches will result in a significant decrease in the minimum *Xcorr* threshold required to achieve a desired FDR: we observed the median *Xcorr* value of rescued PSMs, across all charge states, to be 1.52, with the lowest rescued PSM yielding an *Xcorr* of 0.67, albeit with a *dCn* of 0.08. Legacy SEQUEST users will likely need some adjusting to these unusually low score results, but with careful application of strategies to estimate dataset false discovery rates21 and calculate PSM posterior probability assignments26, we hope they will also ultimately appreciate the benefit of high mass precision to the improved accuracy of their datasets.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

We thank Mike Senko, Justin Blethrow and colleagues at ThermoFisher Scientific for facilitating transfer of the SEQUST source code, additional platform testing, and general discussions. We thank Arminja Kettenbach for preparing the HeLa cell lysate digest and phosphopeptide-enriched sample, and also thank Jason Gilmore for critical reading of the manuscript. This work was supported by the National Institutes of Health grant P20-RR018787 from the IDeA Program of the National Center for Research Resources (to SAG) and a pre-doctoral fellowship from the National Institute of General Medical Sciences grant T32-GM008704 (to BKF).

## References

1. Aebersold R, Mann M. Nature. 2003; 422:198–207. [PubMed: 12634793]
2. Makarov A, Denisov E, Kholomeev A, Balschun W, Lange O, Strupat K, Horning S. Anal Chem. 2006; 78:2113–2120. [PubMed: 16579588]
3. Olsen JV, Schwartz JC, Griep-Raming J, Nielsen ML, Damoc E, Denisov E, Lange O, Remes P, Taylor D, Splendore M, Wouters ER, Senko M, Makarov A, Mann M, Horning S. Mol Cell Proteomics. 2009
4. Second TP, Blethrow JD, Schwartz JC, Merrihew GE, MacCoss MJ, Swaney DL, Russell JD, Coon JJ, Zabrouskov V. Anal Chem. 2009; 81:7757–7765. [PubMed: 19689114]
5. Haas W, Faherty BK, Gerber SA, Elias JE, Beausoleil SA, Bakalarski CE, Li X, Villen J, Gygi SP. Mol Cell Proteomics. 2006
6. Hoopmann MR, Finney GL, MacCoss MJ. Anal Chem. 2007; 79:5620–5632. [PubMed: 17580982]
7. Nesvizhskii AI. Methods Mol Biol. 2007; 367:87–119. [PubMed: 17185772]
8. Kall L, Canterbury JD, Weston J, Noble WS, MacCoss MJ. Nat Methods. 2007; 4:923–925. [PubMed: 17952086]
9. Eng JK, Mccormack AL, Yates JR. Journal of the American Society for Mass Spectrometry. 1994; 5:976–989.
10. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. Electrophoresis. 1999; 20:3551–3567. [PubMed: 10612281]
11. Craig R, Beavis RC. Bioinformatics. 2004; 20:1466–1467. [PubMed: 14976030]
12. Craig R, Beavis RC. Rapid Commun Mass Spectrom. 2003; 17:2310–2316. [PubMed: 14558131]
13. Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, Maynard DM, Yang X, Shi W, Bryant SH. J Proteome Res. 2004; 3:958–964. [PubMed: 15473683]
14. Bakalarski CE, Haas W, Dephoure NE, Gygi SP. Anal Bioanal Chem. 2007; 389:1409–1419. [PubMed: 17874083]
15. Elias JE, Haas W, Faherty BK, Gygi SP. Nat Methods. 2005; 2:667–675. [PubMed: 16118637]
16. Kapp EA, Schutz F, Connolly LM, Chakel JA, Meza JE, Miller CA, Fenyo D, Eng JK, Adkins JN, Omenn GS, Simpson RJ. Proteomics. 2005; 5:3475–3490. [PubMed: 16047398]
17. Sadygov RG, Eng J, Durr E, Saraf A, McDonald H, MacCoss MJ, Yates JR. Journal of Proteome Research. 2002; 1:211–215. [PubMed: 12645897]

18. Park CY, Klammer AA, Kall L, MacCoss MJ, Noble WS. J Proteome Res. 2008; 7:3022–3027. [PubMed: 18505281]

19. Edwards, N.; Lippert, R. Algorithms in Bioinformatics, Proceedings. Guigo, R.; Gusfield, D., editors. Vol. 2452. 2002. p. 68-81.

20. Tabb DL, Narasimhan C, Strader MB, Hettich RL. Anal Chem. 2005; 77:2464–2474. [PubMed: 15828782]

21. Elias JE, Gygi SP. Nat Methods. 2007; 4:207–214. [PubMed: 17327847]

22. Eng JK, Fischer B, Grossmann J, Maccoss MJ. J Proteome Res. 2008; 7:4598–4602. [PubMed: 18774840]

23. Venable JD, Xu T, Cociorva D, Yates JR 3rd. Anal Chem. 2006; 78:1921–1929. [PubMed: 16536429]

24. Beausoleil SA, Jedrychowski M, Schwartz D, Elias JE, Villen J, Li J, Cohn MA, Cantley LC, Gygi SP. Proc Natl Acad Sci U S A. 2004; 101:12130–12135. [PubMed: 15302935]

25. Hsieh EJ, Hoopmann MR, Maclean B, Maccoss MJ. J Proteome Res. 2010; 9:1138–1143. [PubMed: 19938873]

26. Kall L, Storey JD, MacCoss MJ, Noble WS. J Proteome Res. 2008; 7:29–34. [PubMed: 18067246]

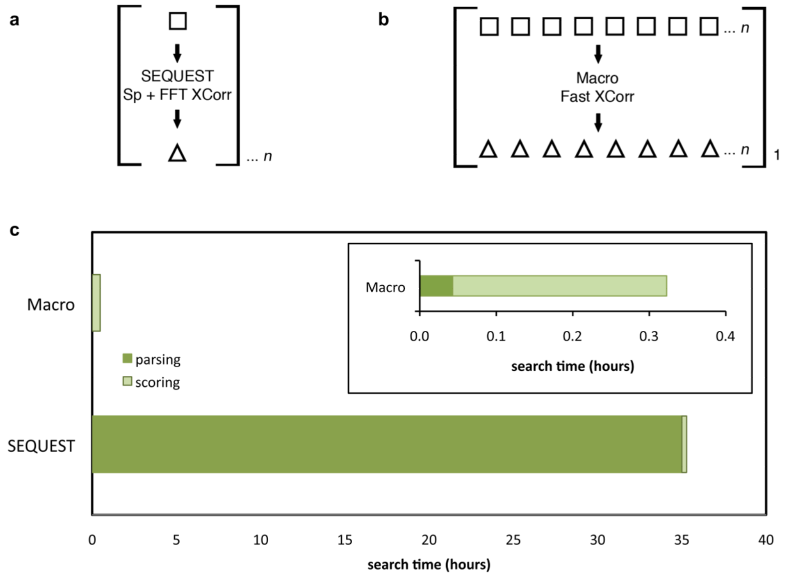

**Figure 1. Comparison of legacy SEQUEST and MacroSEQUEST (Macro) workflows**

**A)** Scheme of the serial nature of SEQUEST use. To analyze *n* number of spectra, SEQUEST must be executed *n* number of times, parsing the database with each instance and using both preliminary and *XCorr* scores. **B)** Scheme of the Macro workflow. Macro considers all spectra in a single analysis, and therefore requires only one pass through a target database during a search. **C)** Analysis of the time distribution for SEQUEST and Macro searches of 10,000 spectra with a target-decoy human (UniProt) database. SEQUEST spends approximately 1% of the search time on scoring functions while Macro spends only 13% of the search time parsing the database. The inset shows an expanded view of the Macro time distribution for the first twenty minutes of the search time.

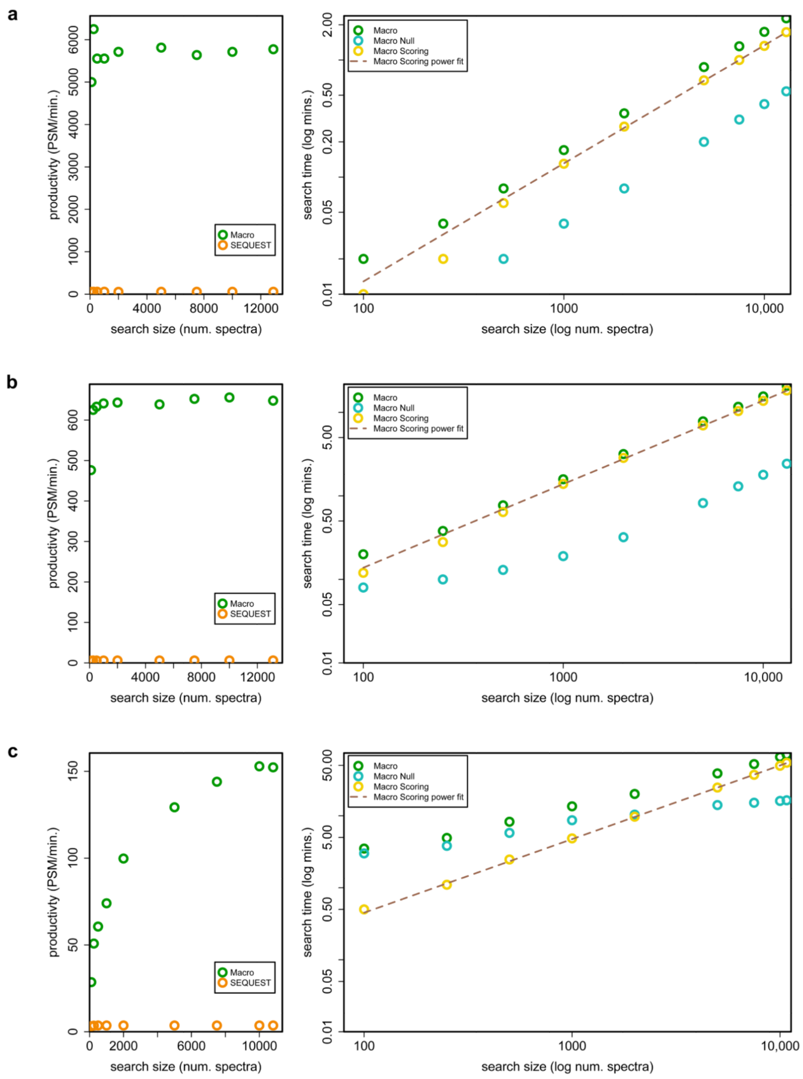

**Figure 2. Productivity of SEQUST and Macro under typical experimental conditions**

We created three biological samples to use in testing the performance of Macro (green circles) versus legacy SEQUEST (orange circles) in common applications: **A)** yeast 80 – 130kDa protein digest, **B)** human 80 – 130kDa protein digest, and **C)** human phosphorylation samples, all analyzed by 90-minute gradient LC-Orbitrap-MS/MS. Variable, random portions of each full collection of MS/MS spectra were searched to define the productivity (number of MS/MS spectra searched/minute of search time) of each search type as shown in the left panel. A version of Macro that does not produce scores and is useful as a measurement of database parsing work, Macro Null, was used to calculate the time spent on scoring by difference. In the right panel, the Macro, Macro Null and calculated Macro Scoring productivity distributions are plotted in log for the variable, random portion search sizes versus search time. A power function was fitted to the Macro Scoring distribution using a non-linear least squares method in R. Note that Macro is 120x, 110x, and 42x faster than SEQUEST under each search condition, respectively.

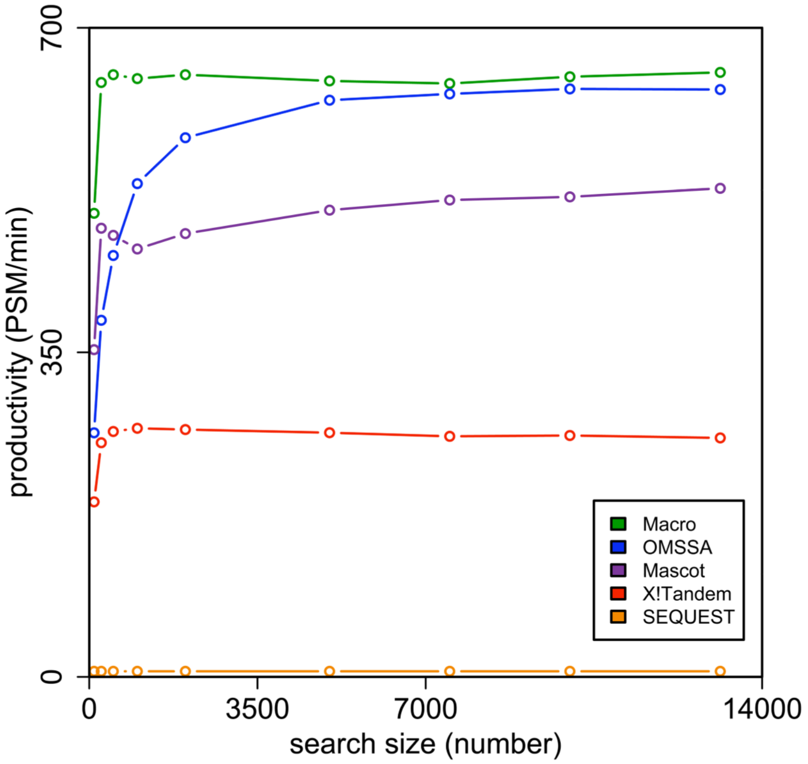

**Figure 3. Comparison of Macro performance with contemporary database searching algorithms**

The same dataset in Figure 2B was re-searched with other commonly used database searching algorithms, including Mascot, OMSSA, and X!Tandem. All algorithms were run in single-threaded mode to provide an accurate comparison.

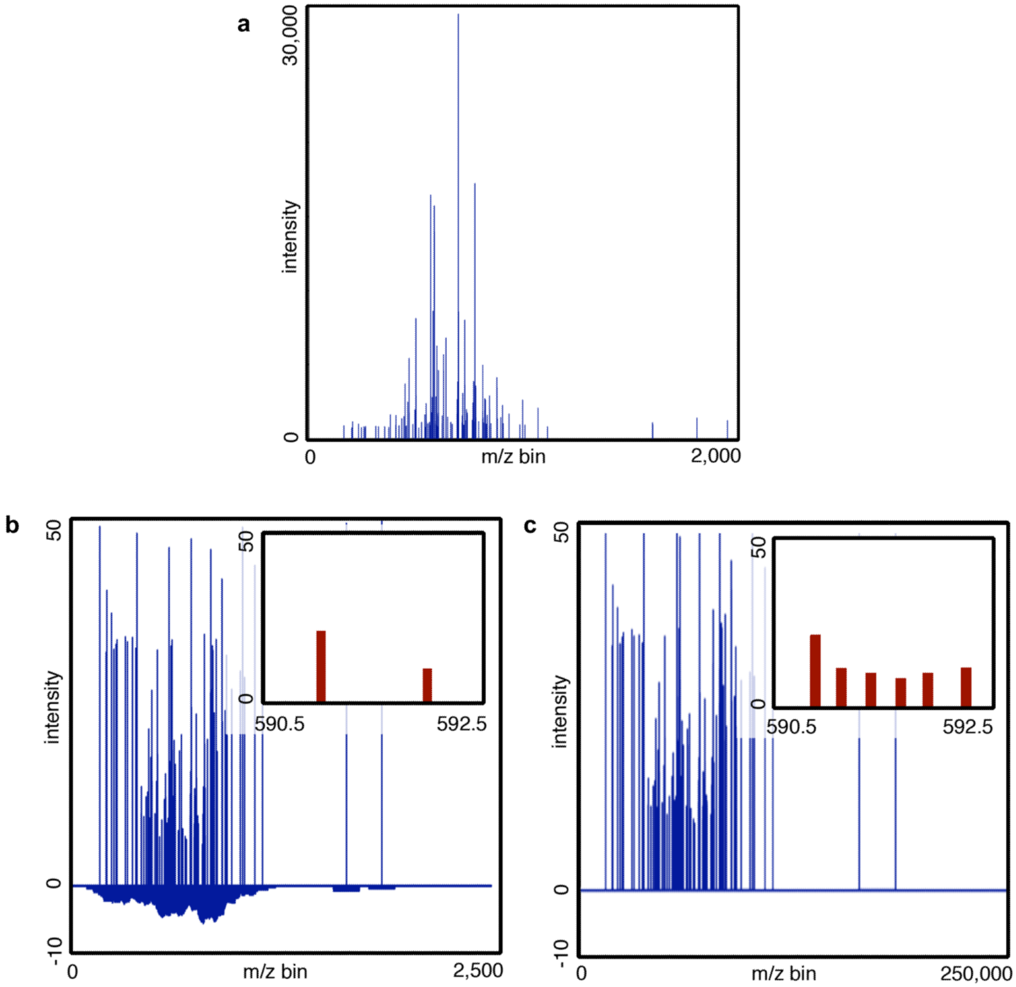

**Figure 4. High resolution MS/MS spectra and Macro**

**A)** A typical, raw high mass accuracy MS/MS spectrum generated using an LTQ-Orbitrap (R = 15,000). **B)** The raw spectrum in **A)** pre-processed for fast *Xcorr* using 1.0 *m/*z-wide bins to define the spectrum array. **C)** The same spectrum pre-processed for fast *Xcorr* using 0.025 *m/z*-wide bins. Note in the inset that multiple ion features persist in a 2 *m/z*-wide space in the higher resolution array, while they are compressed into two features when searches are performed at unit resolution.

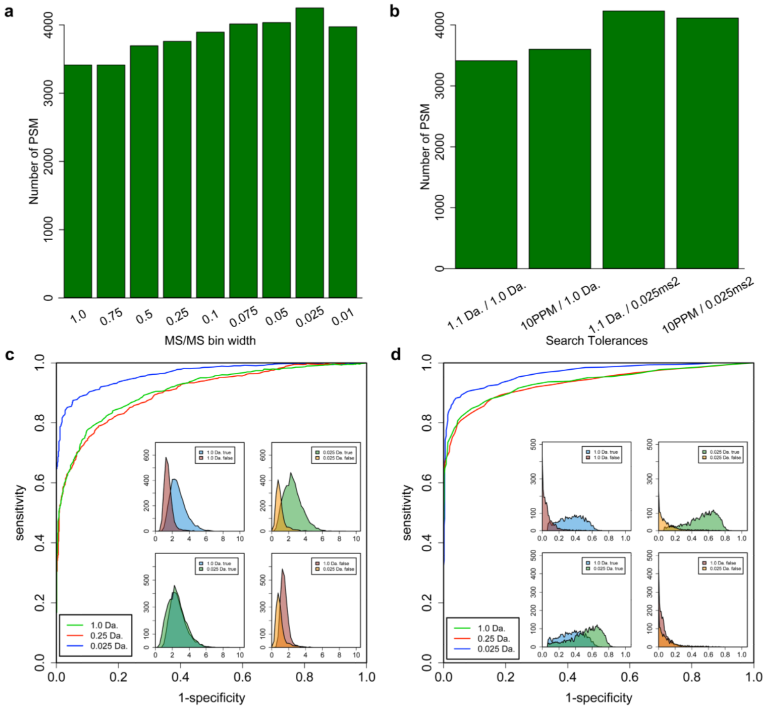

**Figure 5. Behavior of primary scoring metrics for high resolution correlation analysis**

**A)** Histogram depicting the sensitivity of an LC-Orbitrap-MS/MS analysis as a function of fragment ion bin widths at FDR < 1%. **B)** High resolution correlation analysis rescues false negatives that are inaccessible by high mass accuracy precursor ion searching alone. **C)** ROC plot using XCorr filtering during 1.0, 0.25 and 0.025 *m/z*-wide bin searches. Insets show histograms of the behavior of *Xcorr* during 1.0 and 0.025 *m/*z-wide bin searches. The 1.0 *m/z* TP distribution is blue, while the 1.0 *m/z* FP distribution is red. The 0.025 *m/z* TP distribution is green, while the 0.025 *m/z* FP distribution is orange. Note that while the primary score for TPs remains largely unchanged, FP scores are significantly reduced. **D)** ROC plot using dCn filtering during 1.0, 0.25 and 0.025 *m/z*-wide bin searches. Insets show histograms of the behavior of *dCn* during 1.0 and 0.025 *m/z*-wide bin searches, the distributions are colored the same as in **C)**.

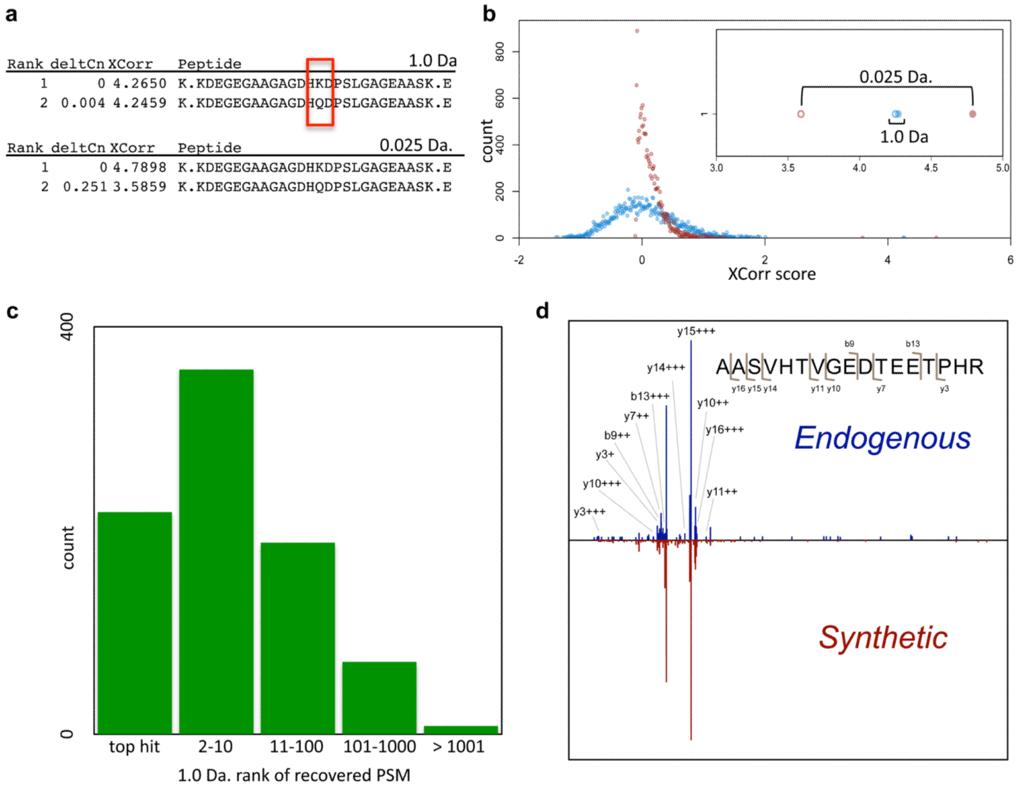

**Figure 6. Characterization of PSMs “rescued” by high resolution correlation analysis**

**A)** Top-ranked PSMs and scores for a given MS/MS search result using 1.0 and 0.025 *m/z*-wide bins. Note that these sequences are identical, except for a K/Q substitution in the middle of the peptide, marked by a red box. While the 1.0 m/z-wide bin search rank 1 PSM satisfies a precursor ion tolerance and *Xcorr* cutoff, the *dCn* (0.005) precludes unambiguous TP assignment. The top-ranked PSMs and scores from the 0.025 *m/z*-wide bin search are shown below. The resolving power of much narrower fragment ion bins clearly allows for discrimination of the K-containing ion fragments, resulting in a *dCn* of 0.36 and an unambiguous TP PSM assignment. **B)** Graphical representation of the distribution of candidate PSMs by *Xcorr* for the 1.0 (blue) and 0.025 (red) *m/z*-wide bin searches of the PSM from **A). C)** Histogram depicting the relative ranks of rescued PSMs in the lower resolution search. The lowest ranked PSM was rescued from the 4,715th position. **D)** Reciprocal MS/MS plots, sequence and matched fragment ions for the PSM rescued from the 4,715th rank position. The upper MS/MS spectrum is from the endogenous peptide from HeLa cell lysate, and the lower MS/MS spectrum is from the synthetic peptide ‘AASVHTVGEDTEETPHR’.