



Published in final edited form as:

*J Proteome Res.* 2011 July 1; 10(7): 2896–2904. doi:10.1021/pr200118r.

## ScanRanker: Quality Assessment of Tandem Mass Spectra via Sequence Tagging

Ze-Qiang Ma<sup>1</sup>, Matthew C. Chambers<sup>1</sup>, Amy-Joan L. Ham<sup>2,3</sup>, Kristin L. Cheek<sup>2,3</sup>, Corbin W. Whitwell<sup>2,3</sup>, Hans-Rudolf Aerni<sup>4,5</sup>, Birgit Schilling<sup>6</sup>, Aaron W. Miller<sup>6</sup>, Richard M. Caprioli<sup>2,4,5</sup>, and David L. Tabb<sup>1,2,3,5,7</sup>

<sup>1</sup>Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, Tennessee 37232-8340

<sup>2</sup>Department of Biochemistry, Vanderbilt University Medical Center, Nashville, Tennessee 37232-0146

<sup>3</sup>Jim Ayers Institute for Precancer Detection and Diagnosis, Vanderbilt-Ingram Cancer Center, Nashville, Tennessee 37232-6350

<sup>4</sup>Department of Chemistry, Vanderbilt University, Nashville, Tennessee 37235

<sup>5</sup>Mass Spectrometry Research Center, Vanderbilt University Medical Center, Nashville, Tennessee 37232-8575

<sup>6</sup>Buck Institute for Age Research, Novato, California 94945

### Abstract

In shotgun proteomics, protein identification by tandem mass spectrometry relies on bioinformatics tools. Despite recent improvements in identification algorithms, a significant number of high quality spectra remain unidentified for various reasons. Here we present ScanRanker, an open-source tool that evaluates the quality of tandem mass spectra via sequence tagging with reliable performance in data from different instruments. The superior performance of ScanRanker enables it not only to find unassigned high quality spectra that evade identification through database search, but also to select spectra for de novo sequencing and cross-linking analysis. In addition, we demonstrate that the distribution of ScanRanker scores predicts the richness of identifiable spectra among multiple LC-MS/MS runs in an experiment, and ScanRanker scores assist the process of peptide assignment validation to increase confident spectrum identifications. The source code and executable versions of ScanRanker are available from <http://fenchurch.mc.vanderbilt.edu>.

### Keywords

spectral quality; sequence tagging; bioinformatics; tandem mass spectrometry; cross-linking

### Introduction

Mass spectrometry (MS)-based proteomics, especially shotgun proteomics, offers a remarkably powerful approach for identification of proteins in complex biological samples. Hundreds of thousands of tandem mass spectra are frequently generated in proteomics experiments, presenting a huge challenge to data analysis. Despite recent improvements in

<sup>7</sup>Corresponding author: (phone) 615-936-0380; (fax) 615-343-8372; david.l.tabb@vanderbilt.edu.

identification methods, a significant number of high quality spectra remain unidentified due to modifications, incompleteness of protein databases, constrained search parameters and the deficiencies of the scoring methods in database search tools. These spectra represent meaningful biological information and are potentially identifiable with alternative approaches<sup>1</sup>, such as blind modification search and de novo sequencing. An automated spectral quality assessment tool helps to ameliorate these problems. It can be used to find unidentified high quality spectra for subsequent analysis and helps to select high quality spectra for de novo sequencing.

Mass spectrometry has become a method of choice to characterize cross-linked proteins<sup>2</sup>. The identification of cross-linked peptides, however, is quite a daunting job due to the overwhelming number of possible matches and the difficulty of interpreting spectra from cross-linked peptides. Although several bioinformatics tools have been developed to relieve these difficulties, manual confirmation of cross-linked peptides is generally necessary<sup>2</sup>. A spectral quality assessment tool could facilitate this process by providing a ranked list of spectra for manual interpretation. High quality spectra with large precursor mass and high charge states are more likely to be derived from cross-linked peptides.

The spectral quality score can also be used in the process of peptide assignment validation. In database search, software tools usually assign different scores to measure the match between spectrum and peptide (e.g. Xcorr from Sequest<sup>3</sup> and IonScore from Mascot<sup>4</sup>), which are subsequently used in statistical analysis to estimate false discovery rates. The spectral quality score could become an additional score in this process, because high quality spectra are more likely to produce confident peptide identifications. Sequence tagging is an approach for peptide identification that infers a partial sequence (a tag) from a spectrum and then evaluates this partial sequence and the rest of the spectrum against candidate peptides from a protein database<sup>5</sup>. The scoring methods in sequence tagging algorithms are applicable for quality assessment of tandem mass spectra. A high quality spectrum of a peptide is expected to contain a series of consecutive fragment ions corresponding to peptide bond breakages<sup>6</sup>. These fragments provide a basis for partial sequence inference that result in multiple tags with good scores. Conversely, if no sequence tags can be inferred from a spectrum, it is unlikely that the spectrum will produce a high score in database search. Sequence tagging is a robust approach for spectral quality assessment because even modified or mutated peptides can produce consecutive fragment ions. Recently, we developed a novel sequence tagging algorithm, DirecTag<sup>7</sup>, which demonstrated superior accuracy in comparison to existing sequence tagging tools. In this work, we explored the use of DirecTag along with other metrics for spectral quality assessment.

Several spectral quality assessment tools have been developed in recent years<sup>8-15</sup>. Pioneering work by Bern et al. predicted spectral quality based on a set of handcrafted features<sup>8</sup>. Other studies by Xu et al.<sup>10</sup> as well as Salmi et al.<sup>11</sup> reported a quadratic discriminant function and a random forest classifier to separate good and bad spectra, respectively. Na et al.<sup>13</sup> proposed a cumulative intensity normalization method for quality assessment, while Flikka et al.<sup>12</sup> tested several machine learning classifiers in data from three different mass spectrometers, recognizing that the performance of classifiers is greatly affected by the type of instrument. More recently, Nesvizhskii et al.<sup>15</sup> developed QualScore, which produces accurate results to find unassigned good spectra after database search. In these prior studies, the proposed methods were usually evaluated based on their performance of removing low quality spectra and recovering unassigned high quality spectra. In fact, quality assessment tools are useful for a wide variety of applications that have not previously been demonstrated. These tools may help to prioritize spectra for de novo sequencing and cross-linking analysis, which are usually very time-consuming processes relying heavily on manual inspection. Besides, since high quality spectra are more likely to

produce confident identifications in database search, the quality assessment tools can also be used for quality control of data sets in large-scale proteomic studies.

In this work, we present ScanRanker, a new software tool that evaluates spectral quality via sequence tagging. We evaluate ScanRanker using a variety of data sets from multiple instrument platforms with different sample complexities. We demonstrate that ScanRanker can be used both to recognize high quality spectra that fail identification and to remove low quality spectra prior to database search. In addition, we demonstrate several applications of spectral quality score that are not explored in existing publications. We show ScanRanker scores can be used to predict the richness of identifiable spectra among LC-MS/MS runs in an experiment. We demonstrate the use of ScanRanker scores in the process of peptide assignment validation. We also demonstrate that ScanRanker helps to select high quality spectra for de novo sequencing and cross-linking analysis.

## Materials and Methods

### ScanRanker Algorithm

ScanRanker makes use of the DirecTag algorithm to infer sequence tags from tandem mass spectra. It then computes a quality score for each spectrum on the basis of three tag-based scoring metrics: “BestTagScore”, “BestTagTIC” and “TagMzRange”. ScanRanker accepts spectra in mzML, mzXML and MGF file formats via use of the ProteoWizard<sup>16</sup> library. Several proprietary formats, such as Thermo RAW files and Bruker YEP files, can also be directly processed with no required installation of vendor-supplied software libraries (a detailed list of supported formats is available at <http://proteowizard.sourceforge.net/docs.html>). ScanRanker can be executed in both Microsoft Windows and Linux systems, though native support for vendor formats requires use of Windows. A graphical user interface (GUI) was created in C# for Windows users and a helper program, IonMatcher, was developed to visualize ScanRanker results and enable interactive manual inspection of peptide-spectrum matches. The screenshots of the ScanRanker and IonMatcher GUI are shown in Supplemental File 1. The source code and executable versions of ScanRanker are available from <http://fenchurch.mc.vanderbilt.edu>.

### BestTagScore Subscore

DirecTag evaluates inferred sequence tags on the basis of peak intensity, m/z fidelity and complementarity<sup>7</sup>. Each tag is assigned a p-value to represent the probability that a better score would have resulted by random chance. Here we make use of the score of the top ranked tag as the “BestTagScore” subscore for spectral quality assessment. Spectra that are capable of generating high quality tags are more likely to be good spectra.

### BestTagTIC Subscore

To infer sequence tags, DirecTag constructs a graph comprising nodes representing peaks and edges representing pairs of peaks that are separated by amino acid masses. DirecTag seeks out consecutive edges in this graph to enumerate sequence tags. For example, a set of four connected nodes in the graph may constitute a tag of three amino acids. Each node in a spectrum graph is associated with a peak intensity value. The “BestTagTIC” subscore sums up peak intensities of the top ranked tag. A high quality spectrum is expected to have a higher “BestTagTIC” subscore than low quality ones in a data set. Spectra that are higher in intensity are more likely to produce tags of high TIC.

### TagMzRange Subscore

Each inferred tag corresponds directly to a series of fragments in a tandem mass spectrum. The “m/z range” of a tag is the m/z distance that extends from the first peak to the last peak

of the tag. By examining all enumerated tags, the “TagMzRange” subscore describes the widest range of  $m/z$  values for a spectrum that is spanned by tags. For a spectrum generating many tags, the “TagMzRange” subscore is equal to the  $m/z$  range between the lowest  $m/z$  peak and the highest  $m/z$  peak across all enumerated tags minus any  $m/z$  areas that are not spanned by tags. If tags can be generated from a wide  $m/z$  range in a spectrum, it is more likely that this spectrum will be identifiable by computational tools.

### Spectral Quality Score

Three subscores are subjected to logarithmic transformation and normalized before generating a final quality score. The normalization of each subscore is performed by subtracting the mean of subscores in that data set, and then divided by the interquartile range of these subscores. Spectra with no inferred tags or the best scored tags exceeded the threshold specified in configuration file, usually 10–20% of spectra in a data set, are considered as low quality spectra and are excluded in the calculation of mean and interquartile range. ScanRanker computes the average of three normalized subscores as the final quality score. Multiple LC-MS/MS runs, such as MudPIT or gel band runs, can be optionally grouped together as a single experiment, for which the mean and interquartile range of subscores across all data sets will be used for normalization.

### Data Sources

The evaluation of the ScanRanker algorithm employed several data sets collected from different instrument platforms (see Table 1). The configurations of ScanRanker and other software tools are given in Supplemental File 1. Full experimental and data processing details of data sets are given in Supplemental File 2. The database search results were processed by IDPicker<sup>17, 18</sup> software for peptide validation and protein assembly. Throughout this study, IDPicker was configured to derive score thresholds to yield a 2% False Discovery Rate (FDR). The data sets are available for download from Vanderbilt University Mass Spectrometry Research Center’s web site (<http://www.mc.vanderbilt.edu/msrc/bioinformatics/data.php>).

## Results and Discussion

To establish ScanRanker’s effectiveness in quality estimation, we first evaluated its three metrics for discrimination. After establishing its scoring discrimination, we tested its real-world performance for recognition of unidentified high quality spectra and prediction of richness of identifiable spectra. We also demonstrated its applications in peptide validation, de novo sequencing and cross-linking analysis. These tests establish ScanRanker as a robust and effective algorithm for spectral quality assessment of data from various instruments in a wide variety of applications.

### Subscore Evaluation

ScanRanker evaluates spectral quality based on “BestTagScore”, “BestTagTIC” and “TagMzRange” subscores. To test the effectiveness of subscores, The “DLD1 LTQ”<sup>18</sup> data set was searched by MyriMatch<sup>19</sup>, Sequest<sup>3</sup> and X!Tandem<sup>20</sup>. The discriminating power of each subscore is illustrated via receiver operating characteristic (ROC) curves in Figure 1. Each subscore may be used to discriminate spectral quality between identified and unidentified spectra. By combining the three subscores, however, ScanRanker achieves better discrimination than by using any single subscore alone. Results obtained after testing any combination of two subscores were exceeded by combining all three subscores (data not shown).

We tested both mean and median for subscore normalization during the development of ScanRanker algorithm, and they worked equally well because of small differences between these values. For example, the average difference between mean and median of “DLD1 LTQ” data set (4 replicates) are 1%, 6% and 2% for “BestTagScore”, “BestTagTIC” and “TagMzRange” subscores, respectively. We choose the mean of subscores for normalization because it is less expensive to compute than the median. More importantly, if ScanRanker scores need to be adjusted across multiple files, the mean of subscores across these files can be easily calculated based on the sum of subscores and the total count of spectra. ScanRanker computes the quality score by averaging three normalized subscores. If the subscores differed considerably in their discriminating powers, simply averaging the subscores would reduce the discriminating power of ScanRanker overall. To test the discrimination difference between optimized score weights and equal weights, each subscore was assigned a weight from 0 to 1 with 0.1 increments, and the summation of weighted subscores was used to calculate the area under ROC curve (AUC). The best possible weighting yielded an AUC less than 1% higher than the equal weight approach. As a result, we opted to use equal weights for simplicity.

### Removal of Low Quality Spectra

Low quality spectra, particularly from ion trap mass spectrometers, often generate a significant amount of computational overhead but contribute little to protein identification. Filtering these spectra via ScanRanker prior to search can save time in identification. To test ScanRanker’s performance in removing low quality spectra, we analyzed three data sets collected from a Thermo Fisher LTQ, an Esquire HCT ultra and a Thermo Fisher LTQ Velos ion trap mass spectrometer. MyriMatch searched these data in two ways: (1) search all spectra, (2) only search the top 60% of high quality spectra as reported by ScanRanker. In all three instruments, more than 94% of the resulting identifications were shared between both searches, and more spectra were identified in the second search than in the first. In the case of the Esquire HCT, almost 5% of the identifications were produced only when the bottom 40% of spectra were pruned away, at the cost of less than 1% of the identifications (see Figure S3A in Supplemental File 3). More identifications were gained by removing low quality spectra prior to database search; low quality spectra are more prone to be matched to decoy sequences, thus increasing the stringency of the threshold applied to all identifications.

Although we retained the top 60% spectra in our test, it should be noted that there is no common threshold that can be applied to all data sets for the selection of high quality spectra. The spectral removal will be more beneficial for large-scale proteomics studies in which multiple biological and technical replicates are analyzed. We recommend determining the percentage of retained spectra by examining the search results of all spectra from a single replicate, then applying the threshold to remove low quality spectra in other replicates. For example, Figure S3B in Supplemental File 3 plots the proportion of retained identified spectra in context of spectra sorted by ScanRanker scores. It is obvious that the top ranked 60% spectra in all three data sets contain more than 95% of identified spectra. Therefore, this threshold could be subsequently used to remove low quality spectra in other replicates before the database search. These figures can be easily generated from ScanRanker output, which comprises a tab-delimited text file including ranked spectra, identification labels and the cumulative sum of identification labels.

### Recovery of Unidentified High Quality Spectra

Simple database search can sometimes fail to identify many spectra that can be identified through additional effort. We employed three publicly available data sets to determine if

ScanRanker scores were predictive of identifications gained through more advanced searching methods.

In the first test, we evaluated the peptides identified through multiple database search algorithms. A single replicate in the “DLD1 LTQ”<sup>18</sup> data set with 12820 MS/MS scans was analyzed using Sequest, yielding 2878 confidently identified spectra. Additional searches using MyriMatch and X!Tandem identified 826 new spectra missed in the Sequest search. All spectra were sorted by ScanRanker scores from high to low quality and were split into deciles. Figure 2A shows the number of initially identified spectra, newly identified spectra and unidentified spectra in each decile. As expected, identified spectra, either by Sequest or additional searches, were associated with higher ScanRanker scores than unidentified spectra.

The second experiment evaluated the peptides gained through semi-tryptic search. For samples dominated by a few major proteins, this strategy improves peptide and protein identification<sup>18</sup>. In this study, we searched the “Serum Orbi”<sup>18</sup> data set using MyriMatch in either fully tryptic or semi-tryptic search mode. Among 6697 MS/MS scans in the data set, 646 spectra were identified in tryptic search, and an additional 928 spectra were generated by semi-tryptic search. Figure 2B plots the distribution of all spectra, split to deciles by ScanRanker scores. It can be observed that the majority of gained spectra by semi-tryptic search were ranked within the top 30% of spectra by ScanRanker.

In the third test, we examined the ability of ScanRanker to find spectra that were unidentified due to modifications and mutations. The “Histone Orbi”<sup>21</sup> data with 9170 MS/MS scans was initially searched using MyriMatch, yielding 641 confidently identified spectra. To find spectra of modified peptides, the data set was searched using TagRecon<sup>22</sup> against a customized database consisting of identified proteins and decoy sequences. TagRecon yielded 672 spectra including common modifications such as acetylation (117 spectra) and deamidation (159 spectra). Among them, 234 spectra were missed in MyriMatch search. Figure 2C shows the distribution of spectra ordered by ScanRanker scores. As in preceding plots, spectra assigned high ScanRanker scores were more likely to be identified through PTM identification software.

### Comparison of ScanRanker to QualScore

QualScore<sup>15</sup> is a tool integrated in the Trans-Proteomic Pipeline that is specifically designed for recognizing spectra that evade identification. We compared the performance of QualScore and ScanRanker on three data sets. To obtain quality scores from QualScore, we analyzed the data sets using Sequest and PeptideProphet<sup>23</sup>, and then processed results using QualScore under the default configuration. Figure 3 shows the ROC curves of ScanRanker and QualScore in three data sets. ScanRanker performed as reliably as QualScore in all tests. ScanRanker displayed slightly better performance than QualScore in the “Histone Orbi” data, possibly because the existence of modified peptides decreased the effectiveness of Sequest/PeptideProphet training, thus diminishing QualScore accuracy. Despite this minor difference, both tools are able to recognize unassigned high quality spectra. QualScore produces accurate results by training its scoring system for each data set based on Sequest/PeptideProphet results, while ScanRanker evaluates spectral quality directly using a sequence tagging approach. Thus, ScanRanker has no dependence on the availability of database search results.

We attempted to include other algorithms in this comparison. Initial tests of msmsEval<sup>14</sup> gave promising discrimination for LTQ data sets, but no training model was provided to enable its use in other types of instruments. The version of the PARC filter<sup>8</sup> that we received from the Yates Laboratory omitted scores for removed spectra, limiting its scope to filtering



spectra prior to database search. In some other tools, the software simply split data sets to “good” and “bad” directories without a report of metrics for each spectrum, limiting conclusions about their scoring discrimination. As a result of these setbacks, we limited our comparison to QualScore.

### Prediction of Richness of Identifiable Spectra

High quality spectra are more likely to be identified in proteomics data analysis. If multiple LC-MS/MS runs are included in an experiment, (for example, MudPIT or 1D gel experiments,) the number of high quality spectra in each data set reveals the richness of identifiable spectra, providing a preliminary overview for the quality of the LC-MS/MS experiment. We sought to demonstrate that the ScanRanker scores are predictive of relative qualities of LC-MS/MS runs in an experiment. Three published data sets, the “MudPIT Orbi”<sup>24</sup>, “IEF Orbi”<sup>24</sup> and “GelBand LTQ”<sup>25</sup> data, were searched using MyriMatch against an IPI Human database. ScanRanker grouped all LC-MS/MS runs in each data set as a single experiment, in which the means and interquartile ranges of subscores across all fractions or gel bands were used for normalization to compute the quality scores. Figure 4 shows the scatter plot between the number of identified spectra in each LC-MS/MS run and the number of retained spectra with ScanRanker scores above different thresholds. Here we used three score thresholds (0, 0.5 and 1). Spectra with score 0 represent scans of better than 60–70% spectra, and spectra scoring 0.5 and 1 have better quality than approximately 85% and 95% of spectra in each experiment, respectively. The distributions of quality scores, however, are dataset-dependent. As expected, the number of high quality spectra predicted by ScanRanker in each data set is highly correlated to the number of identified spectra. For example, a score threshold at 0.5 produced the Pearson correlation coefficients of 0.90, 0.90 and 0.95 for “MudPIT Orbi”, “IEF Orbi” and “GelBand LTQ” data sets, respectively. Therefore, the relative quality of each LC-MS/MS run in an experiment can be estimated by the number of high quality spectra determined by ScanRanker. This is potentially useful for large-scale proteomic studies, in which ScanRanker can be used as a rapid quality control tool to highlight bad LC-MS/MS runs among an experiment.

### Use of Quality Score in Peptide Validation

In proteomics data analysis, database search engines usually generate one or more scores to measure the matches between candidate peptides and experimental spectra. The search results are then processed by either statistical methods (e.g. PeptideProphet) or FDR-based methods (e.g. IDPicker) for peptide validation. In latter methods, usually only scores from database search tools are used to compute FDR. Here we sought to combine spectral quality scores and scores produced by database search tools to increase confident peptide identifications. We searched the “DLD1 LTQ” data using Mascot, Sequest and X!Tandem against an IPI Human database (v3.56). All search results were converted to pepXML files using either an in-house Perl script or software tools in the Trans Proteomics Pipeline. The spectral quality scores generated by ScanRanker were added to pepXML files using a Perl script. IDPicker subsequently read these scores along with search engine scores during peptide validation. The software combined multiple scores by optimizing score weights through a Monte Carlo method<sup>18</sup>, generating a single score for each peptide-spectrum match. In this test, we configured IDPicker to use either the primary scores from a database search tool or these scores plus the spectral quality score. Figure 5 shows the percent overlap of confident spectrum identifications in both settings. Adding spectral quality scores in peptide validation consistently yielded more confident spectrum identifications than using a single score. Mascot benefited significantly more from score combination than Sequest and X!Tandem. Some spectra may be identified only when using the primary score. These spectra, however, are usually less confident identifications that are assigned marginal match scores in database search.

## Selection of Spectra for De Novo Sequencing

De novo sequencing is an alternative, database-independent approach for peptide identification. However, inferring peptides from spectra is a time-consuming process. In this study, for example, PepNovo<sup>26, 27</sup> (release 20100225) took about 8 hours to infer sequences of an Orbitrap data set with 14217 MS/MS scans on a Dell Optiplex 745 computer with an Intel Core 2 Duo 6400 processor and 3 GB of RAM, while ScanRanker only required 3 minutes for spectral quality assessment. Therefore, de novo sequencing could benefit from the application of spectral quality assessment tools by selecting high quality spectra for analysis.

As a state-of-the-art de novo sequencing tool, PepNovo assigns a score to each inferred peptide sequence to evaluate how well it explains the peak pattern in a spectrum. The higher a PepNovo score, the better an inferred peptide matches a spectrum. We employed three data sets to demonstrate that high ScanRanker scores are predictive of high PepNovo scores. The initial comparison of these scores analyzed the “Yeast Velos” data set, in which peptide identification was straightforward. Figure 6A shows the scatter plot between the PepNovo score of the top ranked peptide sequence for each spectrum and its ScanRanker score. ScanRanker scores are highly correlated to PepNovo scores, producing a Pearson correlation coefficient of 0.82. As expected, spectra identified by MyriMatch search tend to associate with high ScanRanker and PepNovo scores.

Next, we evaluated ScanRanker on data sets for which de novo sequencing would be necessary. The “Tardigrade QSTAR” data set is an LC-MS/MS experiment from a 1D gel band from a species of microscopic animals for which genome sequence is unavailable. MyriMatch attempted to produce identifications in a customized database containing proteins of three species that are taxonomically similar to tardigrade (*Drosophila melanogaster* (DROME), *Anopheles gambiae* (African malaria mosquito, ANOGA) and *Caenorhabditis elegans* (CAEEL)). Only spectra for peptides of highly similar proteins would be identified by this approach; only 66 spectra were identified among the 837 MS/MS scans in the set. Figure 6B superimposes these identifications on the scatter plot of PepNovo and ScanRanker scores. PepNovo and ScanRanker both report that many spectra were of high quality and yet failed identification. Pearson correlation between the two algorithms produced a coefficient of 0.72.

Considerable controversy has accompanied the recent publication of proteomics data for fossilized specimens<sup>28</sup>. We sought to characterize the recent “Hadrosaur Orbi” data set to evaluate the inherent identifiability of spectra for these spectra. We began with a database search against a lizard (*Anolis carolinensis*) database, AnoCar1.0, produced by the Broad Institute at MIT and Harvard (<http://www.broadinstitute.org/models/anole>). The result included 189 confidently identified tandem mass spectra, but all matched to keratin or trypsin sequences (our database did not include the chicken sequences employed by the Asara group). We plotted spectra against the corresponding PepNovo and ScanRanker scores (see Figure 6C). Five collagen spectra from the original Asara publication were assigned high ScanRanker quality scores of 1.13, 0.99, 0.97, 1.01 and 1.70; we were unable to match the sixth identification to the corresponding MS/MS spectrum. The hadrosaur data produced the lowest correlation between PepNovo and ScanRanker (0.34), where the best correspondence could be observed in the high scoring domains for the two algorithms. It becomes clear that the data of the “Hadrosaur Orbi” set were disproportionately likely to produce PepNovo scores below zero, suggesting that a large fraction of spectra from this data set could not support confident sequence identifications even if appropriate sequences were available in FASTA.



## Use of ScanRanker in Cross-linking Analysis

Identification of cross-linked peptides by mass spectrometry is a challenging task, mainly because of the high complexity and often low signal intensity in these spectra. Even with the availability of advanced computational tools, manual interpretation or confirmation of cross-linked peptides is generally necessary. Here we sought to demonstrate that ScanRanker helps to prioritize spectra for manual inspection. The published “Crosslink Orbi”<sup>29</sup> data set consists of 1161 MS/MS spectra collected on an LTQ-Orbitrap XL with an ETD module installed (Thermo Scientific). Spectra in quadruply charged or higher charge states were selected for ETD fragmentation to characterize chemically cross-linked GroEL-GroES chaperonin complex. Protein Prospector<sup>30</sup> identified 55 spectra of cross-linked peptides (manually confirmed) and 91 spectra of single peptides. Figure 7 shows the distribution of these spectra, split to deciles by ScanRanker scores. The spectra of cross-linked peptides were associated with high ScanRanker scores, suggesting that ScanRanker is capable of recognizing these spectra, though they are more complicated than spectra of single peptides. The results also indicate that ScanRanker performs well for spectra from ETD fragmentation.

Some spectra were assigned high quality scores but remained unidentified. A manual inspection of these spectra implies that they are likely produced by peptides rather than non-peptide contaminants. These spectra usually contain a large number of peaks. For example, the top 10% of spectra by ScanRanker includes 70 unidentified spectra. The average number of peaks in these spectra is 228, which is much higher than that number of all spectra (91 peaks) in the data set.

## Conclusion

We present a method that assesses quality of tandem mass spectra through sequence tagging. ScanRanker does not require training for each type of data from different mass spectrometers, broadening its use to lab researchers lacking prior experience in statistical learning. In this study, we employed a variety of data sets to demonstrate the effectiveness of ScanRanker for recovery of unidentified high quality spectra and removal of low quality spectra. We showed that ScanRanker can be used to predict the richness of identifiable spectra in LC-MS/MS experiments and to improve peptide validation. We also demonstrate the application of our method to rank spectra for de novo sequencing and cross-linking analysis. The superior performance of ScanRanker established it as a robust and reliable spectral quality assessment tool.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

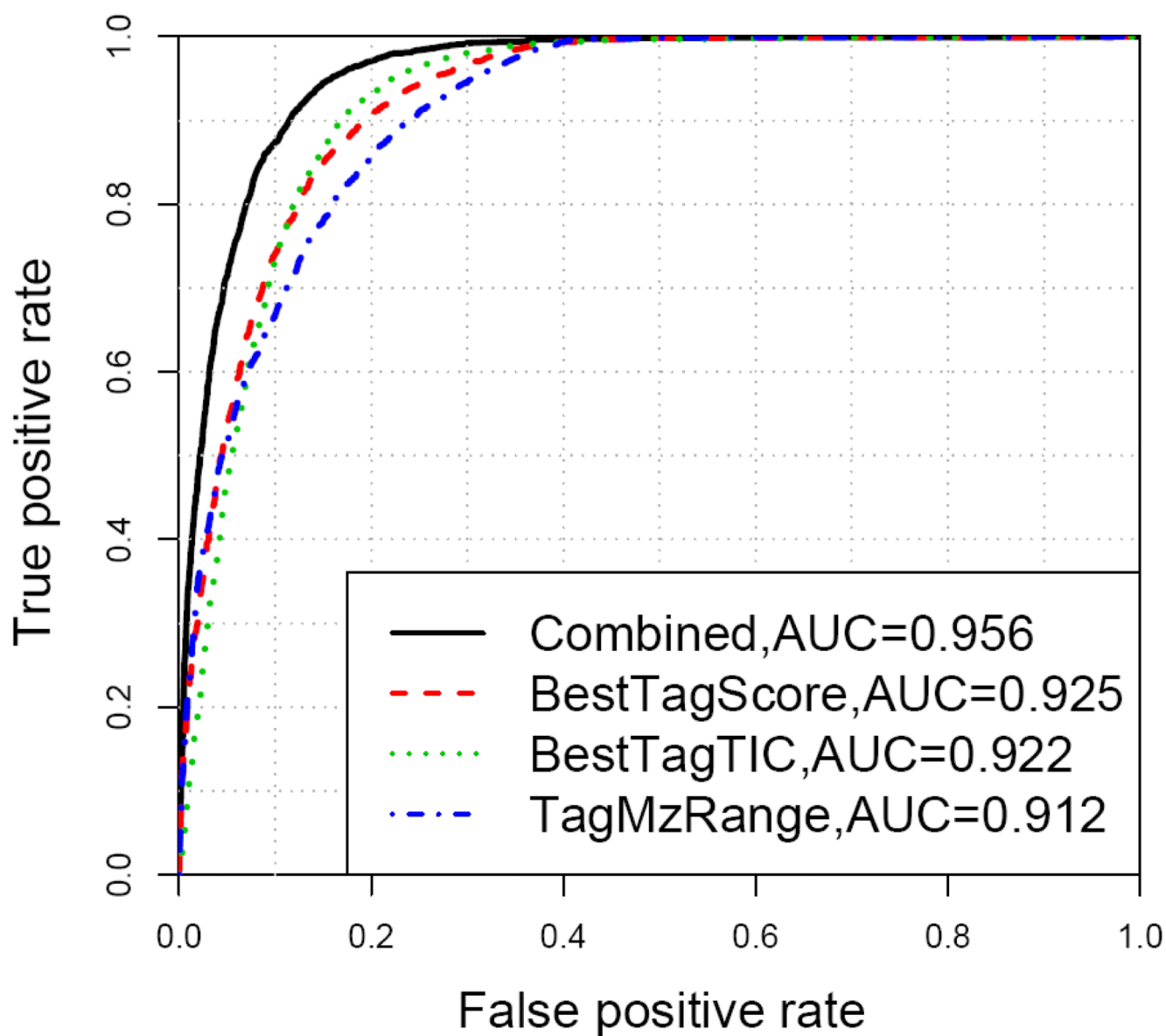
## Acknowledgments

D.L. Tabb, Z.-Q. Ma, and M.C. Chambers were supported by NIH grants R01 CA126218 and U24 CA126479. We would like to thank Robert Chalkley at University of California, San Francisco for providing the “Crosslink Orbi” data set and informing spectra of cross-linked peptides. The “Yeast Velos” data was collected by K. Cheek under NIH grant U24 CA126479 at Vanderbilt University. B. Schilling and A.W. Miller at the Buck Institute for Age Research provided the “Tardigrade QSTAR” data, which was supported by two NIH grants to the Buck Institute for Aging Research that were components of a larger U54 award on Geroscience; UL1 DE019608 (R. E. Hughes), PL1 AG032118 (B. W. Gibson), and a shared instrumentation grant for the QSTAR Elite (NCRR 1S10RR024615 to B. W. Gibson). We would like to thank Tao Xu from Yates Laboratory at the Scripps Research Institute for providing the PARC filter. We thank Surendra Dasari and Lorenzo J Vega-Montoto at Vanderbilt University for valuable discussions.

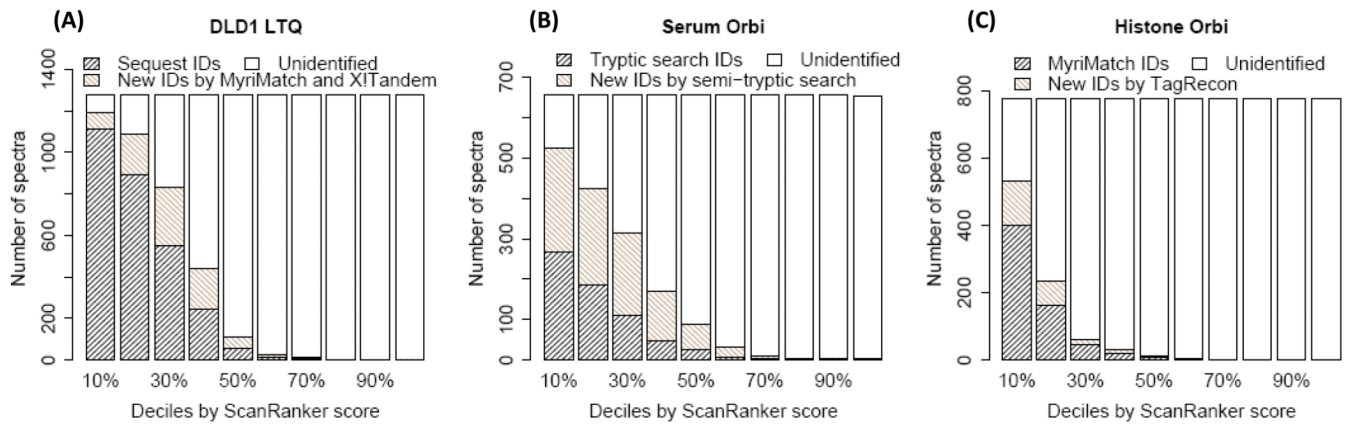
## References

1. Ning K, Fermin D, Nesvizhskii AI. Computational analysis of unassigned high-quality MS/MS spectra in proteomic data sets. *Proteomics*. 2010; 10(14):2712–2718. [PubMed: 20455209]
2. Leitner A, Walzthoeni T, Kahraman A, Herzog F, Rinner O, Beck M, Aebersold R. Probing native protein structures by chemical cross-linking, mass spectrometry, and bioinformatics. *Mol Cell Proteomics*. 2010; 9(8):1634–1649. [PubMed: 20360032]
3. Eng JK, Mann M, Yates JR. An approach to correlate tandem mass-spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* 1994; 5:976–989.
4. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*. 1999; 20(18):3551–3567. [PubMed: 10612281]
5. Mann M, Wilm M. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal Chem*. 1994; 66(24):4390–4399. [PubMed: 7847635]
6. Tabb DL, Friedman DB, Ham AJ. Verification of automated peptide identifications from proteomic tandem mass spectra. *Nat Protoc*. 2006; 1(5):2213–2222. [PubMed: 17406459]
7. Tabb DL, Ma ZQ, Martin DB, Ham AJ, Chambers MC. DirecTag: accurate sequence tags from peptide MS/MS through statistical scoring. *J Proteome Res*. 2008; 7(9):3838–3846. [PubMed: 18630943]
8. Bern M, Goldberg D, McDonald WH, Yates JR 3rd. Automatic quality assessment of peptide tandem mass spectra. *Bioinformatics*. 2004; 20 Suppl 1:i49–i54. [PubMed: 15262780]
9. Purvine S, Kolker N, Kolker E. Spectral quality assessment for high-throughput tandem mass spectrometry proteomics. *Omics*. 2004; 8(3):255–265. [PubMed: 15669717]
10. Xu M, Geer LY, Bryant SH, Roth JS, Kowalak JA, Maynard DM, Markey SP. Assessing data quality of peptide mass spectra obtained by quadrupole ion trap mass spectrometry. *J Proteome Res*. 2005; 4(2):300–305. [PubMed: 15822904]
11. Salmi J, Moulder R, Filen JJ, Nevalainen OS, Nyman TA, Laheesmaa R, Aittokallio T. Quality classification of tandem mass spectrometry data. *Bioinformatics*. 2006; 22(4):400–406. [PubMed: 16352652]
12. Flikka K, Martens L, Vandekerckhove J, Gevaert K, Eidhammer I. Improving the reliability and throughput of mass spectrometry-based proteomics by spectrum quality filtering. *Proteomics*. 2006; 6(7):2086–2094. [PubMed: 16518876]
13. Na S, Paek E. Quality assessment of tandem mass spectra based on cumulative intensity normalization. *J Proteome Res*. 2006; 5(12):3241–3248. [PubMed: 17137325]
14. Wong JW, Sullivan MJ, Cartwright HM, Cagney G. msmsEval: tandem mass spectral quality assignment for high-throughput proteomics. *BMC Bioinformatics*. 2007; 8:51. [PubMed: 17291342]
15. Nesvizhskii AI, Roos FF, Grossmann J, Vogelzang M, Eddes JS, Gruissem W, Baginsky S, Aebersold R. 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*. 2006; 5(4):652–670. [PubMed: 16352522]
16. Kessner D, Chambers M, Burke R, Agus D, Mallick P. ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics*. 2008; 24(21):2534–2536. [PubMed: 18606607]
17. Zhang B, Chambers MC, Tabb DL. Proteomic parsimony through bipartite graph analysis improves accuracy and transparency. *J Proteome Res*. 2007; 6(9):3549–3557. [PubMed: 17676885]
18. Ma ZQ, Dasari S, Chambers MC, Litton MD, Sobecki SM, Zimmerman LJ, Halvey PJ, Schilling B, Drake PM, Gibson BW, Tabb DL. IDPicker 2.0: Improved protein assembly with high discrimination peptide identification filtering. *J Proteome Res*. 2009; 8(8):3872–3881. [PubMed: 19522537]

19. Tabb DL, Fernando CG, Chambers MC. MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *J Proteome Res.* 2007; 6(2):654–661. [PubMed: 17269722]
20. Craig R, Beavis RC. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics.* 2004; 20(9):1466–1467. [PubMed: 14976030]
21. Loecken EM, Dasari S, Hill S, Tabb DL, Guengerich FP. The bis-electrophile diepoxybutane cross-links DNA to human histones but does not result in enhanced mutagenesis in recombinant systems. *Chem Res Toxicol.* 2009; 22(6):1069–1076. [PubMed: 19364102]
22. Dasari S, Chambers MC, Slebos RJ, Zimmerman L, Ham AJ, Tabb DL. TagRecon: high-throughput mutation identification through sequence tagging. *J Proteome Res.*
23. Keller A, Nesvizhskii AI, Kolker E, Aebersold R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal Chem.* 2002; 74(20):5383–5392. [PubMed: 12403597]
24. Slebos RJ, Brock JW, Winters NF, Stuart SR, Martinez MA, Li M, Chambers MC, Zimmerman LJ, Ham AJ, Tabb DL, Liebler DC. Evaluation of strong cation exchange versus isoelectric focusing of peptides for multidimensional liquid chromatography-tandem mass spectrometry. *J Proteome Res.* 2008; 7(12):5286–5294. [PubMed: 18939861]
25. Burgess EF, Ham AJ, Tabb DL, Billheimer D, Roth BJ, Chang SS, Cookson MS, Hinton TJ, Cheek KL, Hill S, Pietenpol JA. Prostate cancer serum biomarker discovery through proteomic analysis of alpha-2 macroglobulin protein complexes. *Proteomics Clin Appl.* 2008; 2(9):1223. [PubMed: 20107526]
26. Frank A, Pevzner P. PepNovo: de novo peptide sequencing via probabilistic network modeling. *Anal Chem.* 2005; 77(4):964–973. [PubMed: 15858974]
27. Frank AM. A ranking-based scoring function for peptide-spectrum matches. *J Proteome Res.* 2009; 8(5):2241–2252. [PubMed: 19231891]
28. Schweitzer MH, Zheng W, Organ CL, Avci R, Suo Z, Freimark LM, Lebleu VS, Duncan MB, Vander Heiden MG, Neveu JM, Lane WS, Cottrell JS, Horner JR, Cantley LC, Kalluri R, Asara JM. Biomolecular characterization and protein sequences of the Campanian hadrosaur *B. canadensis*. *Science.* 2009; 324(5927):626–631. [PubMed: 19407199]
29. Trnka MJ, Burlingame AL. Topographic studies of the GroEL-GroES chaperonin complex by chemical cross-linking using diformyl ethynylbenzene: the power of high resolution electron transfer dissociation for determination of both peptide sequences and their attachment sites. *Mol Cell Proteomics.* 2010; 9(10):2306–2317. [PubMed: 20813910]
30. Chu F, Baker PR, Burlingame AL, Chalkley RJ. Finding chimeras: a bioinformatics strategy for identification of cross-linked peptides. *Mol Cell Proteomics.* 2010; 9(1):25–31. [PubMed: 19809093]



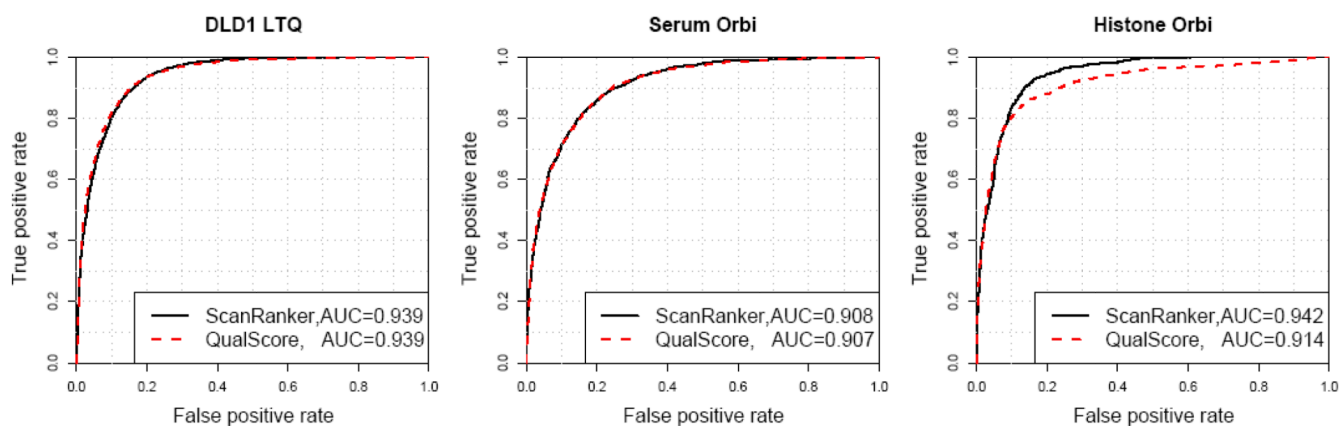
**Figure 1.** Combining three subscores improves the discriminating power of ScanRanker. Tests on the “DL1 LTQ” data set revealed different discrimination in ScanRanker’s subscores. The ROC curves display true positive rate (a.k.a. sensitivity) and false positive rate (a.k.a. 1-specificity) of ScanRanker’s subscores and the combined score. The AUC values show that combining three subscores yields better discrimination than using any single subscore.



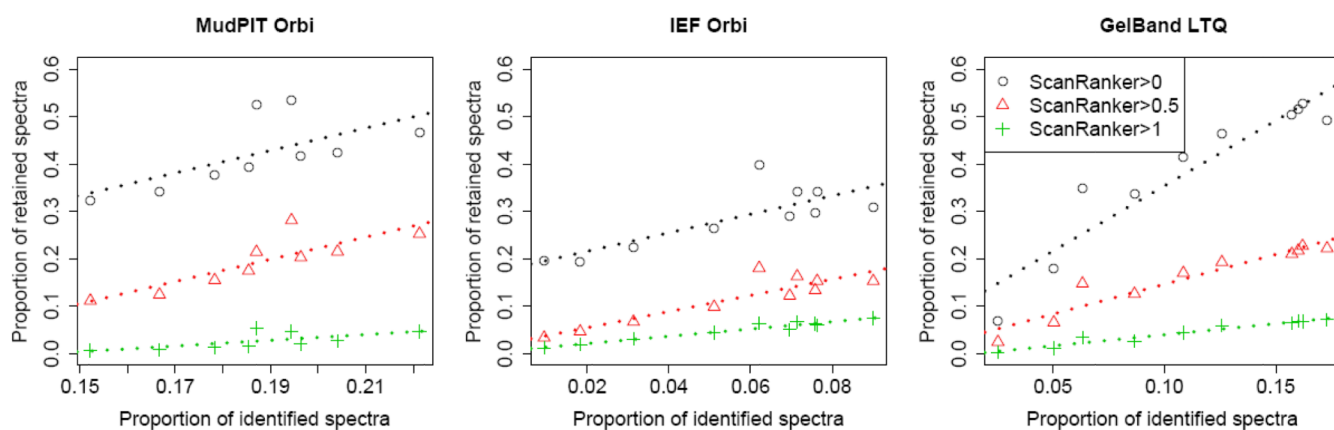
**Figure 2.**

Evaluation of ScanRanker to recover unidentified high quality spectra. Three data sets were reanalyzed by additional search methods to find high quality spectra that were unidentified in initial database searches. Each test represents a typical reason that high quality spectra may be left unidentified in an initial search. (A) The “DLD1 LTQ” data set was initially identified by Sequest search. New identifications (IDs) were added by MyriMatch and X! Tandem searches. (B) The “Serum Orbi” data was searched by MyriMatch in either tryptic or semi-tryptic mode. (C) The “Histone Orbi” data was searched by MyriMatch. A subsequent TagRecon search was performed to identify spectra of mutated or modified peptides. These graphs plot the distributions of initial identifications, new identifications by additional searches and unidentified spectra in deciles by ScanRanker scores. In each panel, the left side represents spectra assigned high ScanRanker quality scores and the right side is low quality spectra. Newly identified spectra tend to associate with better ScanRanker scores in all data sets.



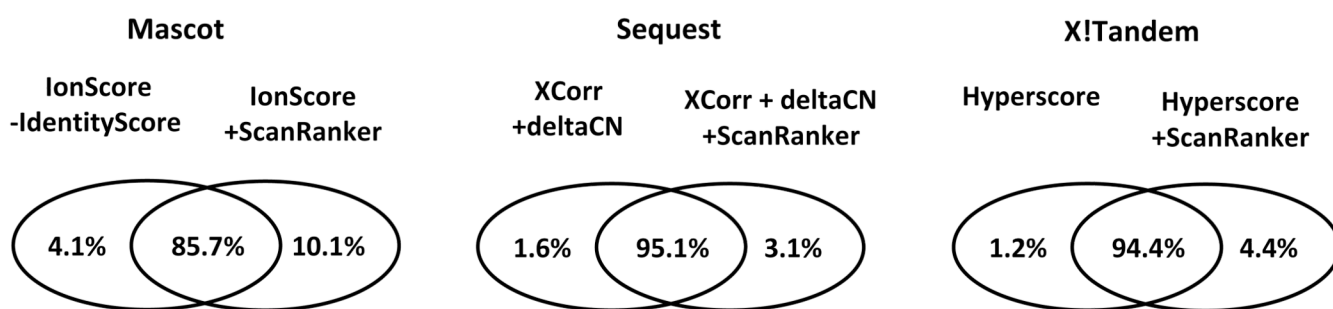


**Figure 3.** Comparison of ScanRanker to QualScore. Spectra in three data sets were separately processed by ScanRanker and QualScore to generate quality scores. ScanRanker performs as well as QualScore in all data sets but does not require Sequest/PeptideProphet analysis for spectral quality assessment.



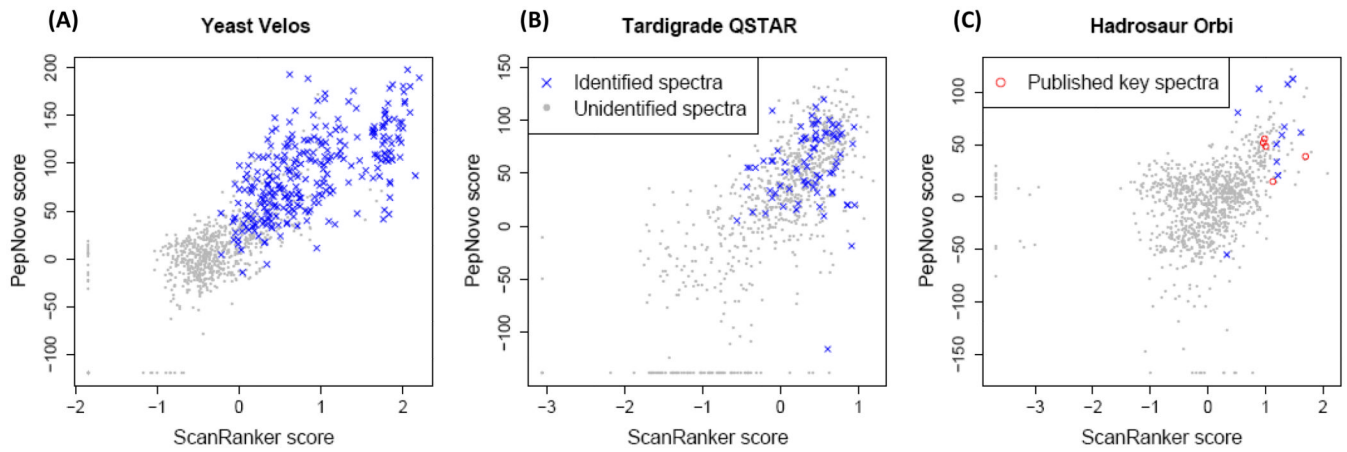
**Figure 4.**

ScanRanker scores predict the richness of identifiable spectra. Each point in the figure represents a single LC-MS/MS run and the dotted lines show the least squares fit of the data. Three ScanRanker thresholds were used to count retained spectra. 9 of 10 LC-MS/MS runs in the MudPIT data set are plotted because the first fraction of the MudPIT experiment generated only 21 spectrum identifications. Each LC-MS/MS run in all three data sets includes about 10000 MS/MS spectra, while the number of identified spectra varies dramatically. The number of spectra assigned high ScanRanker scores correlate to the number of identified spectra, providing relative quality assessment of LC-MS/MS runs in an experiment.



**Figure 5.**

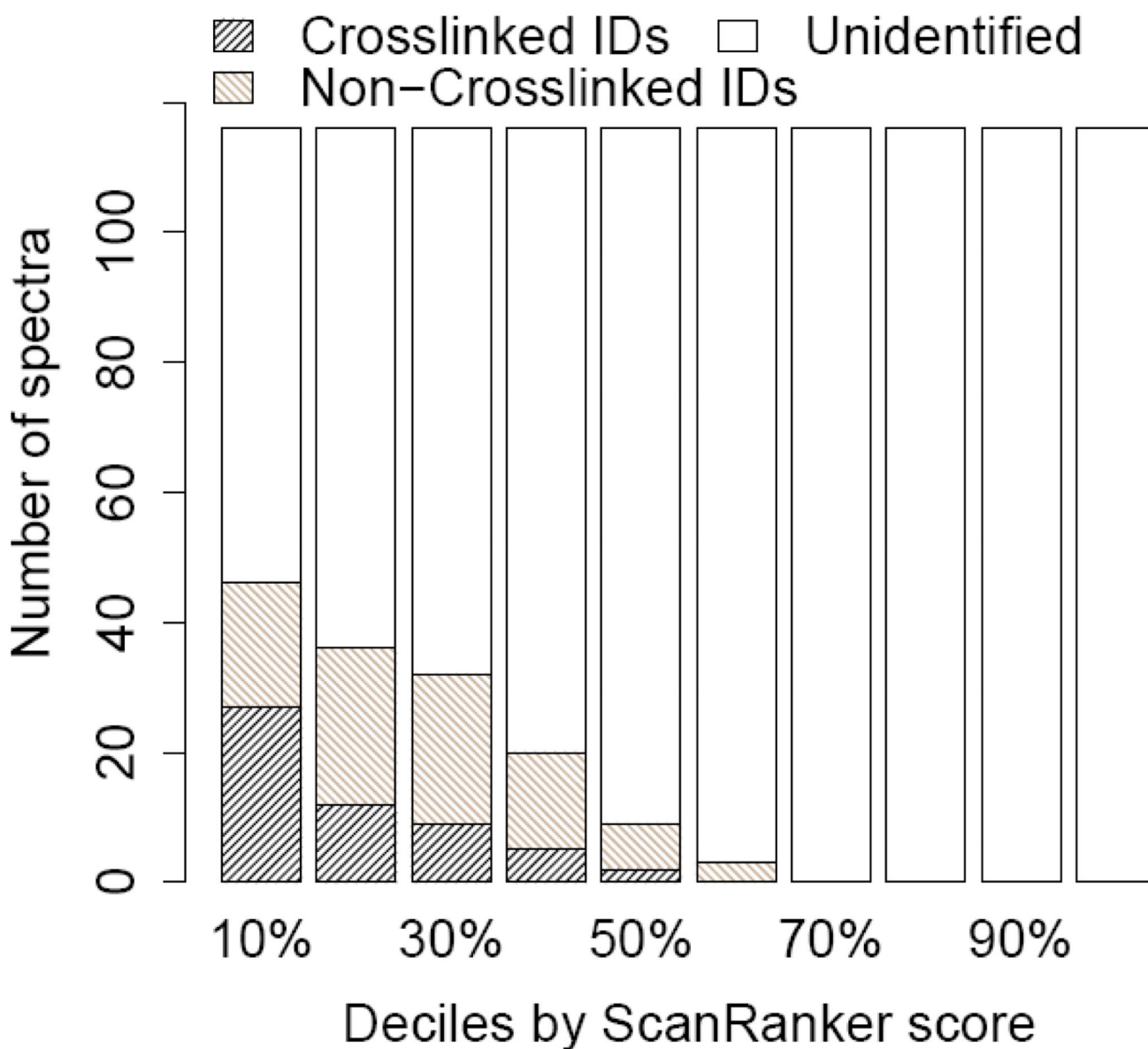
Adding ScanRanker scores in peptide validation increases the number of confident spectrum identifications. “DLD1 LTQ” data set was separately searched by Mascot, Sequest and X! Tandem. ScanRanker scores were added to pepXML files to allow score combination in IDPicker. Mascot scores were combined using either static weights as “IonScore-IdentityScore” or optimized weights as “IonScore + ScanRanker”. Sequest and X!Tandem results were combined by enabling score weights optimization in IDPicker. The Venn diagrams show the percent overlap of identified spectra when using either a single score or combination of two scores. The latter method yielded more spectrum identifications for all searches.



**Figure 6.**

ScanRanker scores can be used to predict de novo sequencing success. Spectra in three data sets were separately processed by ScanRanker and PepNovo. Identifications were generated by searching the spectra using MyriMatch. For clarity, only 1000 spectra were randomly sampled and displayed. When PepNovo reported no peptide for a spectrum, it was visualized as matching the minimum score reported by the software for that data set. Panel C highlights five published key spectra from the Asara group publication. In all three tests, spectra with high ScanRanker scores tend to be assigned high PepNovo scores, implying that ScanRanker can be used to select high quality spectra for de novo sequencing.

## Crosslink Orbi



**Figure 7.** ScanRanker helps to prioritize spectra for manual inspection in cross-linking analysis. The “Crosslink Orbi” data set was processed using Protein Prospector to identify crosslinked and non-crosslinked spectra. The figure plots the distribution of these spectra in deciles by ScanRanker scores. The identified spectra, either crosslinked or non-crosslinked, were associated with high ScanRanker scores, implying that ScanRanker can be used to facilitate cross-linking analysis by ranking spectra for manual inspection.



**Table 1**

Experimental Data Sets Summary. Full experimental and data processing details of all data sets are given in Supplemental File 2.

dataset name	# of files	(average) # of MS/MS scans	identification methods	databases used for search
<i>Removal of Low Quality Spectra</i>				
DLD1 LTQ	4	12913	MyriMatch, Sequest, X!Tandem	IPI.HUMAN.v3.56
Mouse HCT	4	5408	MyriMatch, Sequest, X!Tandem	IPI.MOUSE.v3.62
Yeast Velos	5	38466	MyriMatch, Sequest, X!Tandem	SGD.orf_trans_all.20090303
<i>Recovery of Unidentified High Quality Spectra</i>				
DLD1 LTQ	1	12820	Sequest/MyriMatch, X!Tandem	IPI.HUMAN.v3.56
Serum Orbi	1	6697	MyriMatch, tryptic/semi-tryptic	IPI.HUMAN.v3.56
Histone Orbi	1	9170	MyriMatch/TagRecon	IPI.HUMAN.v3.68
<i>Prediction of Richness of Identifiable Spectra</i>				
MudPIT Orbi	10	9828	MyriMatch	IPI.HUMAN.v3.56
IEF Orbi	10	10897	MyriMatch	IPI.HUMAN.v3.56
GelBand LTQ	10	9520	MyriMatch	IPI.HUMAN.v3.47
<i>Use of Quality Score in Peptide Validation</i>				
DLD1 LTQ	4	12913	Mascot, Sequest, X!Tandem	IPI.HUMAN.v3.56
<i>Selection of Spectra for De Novo Sequencing</i>				
Yeast Velos	1	38560	PepNovo, MyriMatch	SGD.orf_trans_all.20090303
Tardigrade QSTAR	1	837	PepNovo, MyriMatch	SwissProt.DROME.ANOGA.CAEEL.rel56.8
Hadrosaur Orbi	1	14217	PepNovo, MyriMatch	AnoCar1.0
<i>Use of ScanRanker in Cross-linking Analysis</i>				
Crosslink Orbi	1	1161	Protein Prospector	SwissProt.ECOLI.20100810