Analysis of False Discovery Rate Strategies in Shotgun Proteomics

Matteo Venturini (2469579)
Uliana Elizarova (2467726)

December 18, 2025

Abstract

Shotgun proteomics relies on accurate peptide-spectrum matching (PSM) to identify proteins from tandem mass spectrometry data. False Discovery Rate (FDR) estimation is critical for ensuring the reliability of these identifications. This study evaluates the impact of two database search strategies, emulated concatenated and un-concatenated, on PSM confidence at a fixed 1% FDR threshold. Using a Saccharomyces cerevisiae dataset from the ABRF Proteome Informatics Research Group (iPRG) 2015 study, we demonstrate that the concatenated approach, which incorporates a competition model between target and decoy matches, identifies 1.25 times more confident PSMs compared to the conservative un-concatenated strategy. Analysis of score distributions reveals a bimodal pattern for target matches, contrasting with the unimodal decoy distribution, and highlights the context-dependent nature of PSM confidence. Local FDR assessment further emphasizes the need for per-spectrum evaluation, as high-scoring matches exhibit clear separation from noise, while lower-scoring PSMs show increased ambiguity. These findings support the adoption of concatenated search strategies and local error rate methodologies to improve statistical power and confidence in shotgun proteomics workflows.

Keywords: Shotgun proteomics, False Discovery Rate (FDR), peptide-spectrum matching (PSM), concatenated database search, target-decoy competition, local FDR, mass spectrometry, Saccharomyces cerevisiae

Contents

1 Introduction

1.1 Project Objective

The goal of this project is to conduct a detailed investigation into the statistical methods used for error rate control in shotgun proteomics [1]. Specifically, this report examines the effect of different database search strategies on the final list of identified peptide-spectrum matches, or PSMs. The primary comparison is between an un-concatenated database search, where target and decoy databases are searched separately, and an emulated concatenated search, where a competition model is applied to the separate search results. The analysis aims to provide a well-argued recommendation on best practices for setting up database search algorithms by manually calculating and comparing False Discovery Rates, or FDR [2], for each strategy.

1.2 Data Source and Software

The experimental data originates from the ABRF Proteome Informatics Research Group 2015 study [3], publicly available through the ProteomeXchange consortium with the dataset identifier PXD015300 [4]. The specific file analyzed is JD_06232014_sample1-A.raw, which contains tandem mass spectrometry data from a complex sample of Saccharomyces cerevisiae tryptic digest spiked with six standard proteins. The protein sequence database, iPRG2015.fasta, was also obtained from the study's repository. The computational analysis was performed using SearchGUI version 4.3.17 [5], which utilized the Comet search engine for peptide identification [6]. Subsequent data processing, statistical analysis, and visualization were conducted in the R programming language [7], utilizing packages including mzR for parsing search results [8], dplyr for data manipulation [9], and ggplot2 for generating plots [10].

2 Methods

2.1 Protein Sequence Database Preparation

The analysis began with the iPRG2015.fasta file provided by the study authors. This file contains the protein sequences for the Saccharomyces cerevisiae background proteome as well as the six spiked-in standard proteins.

2.2 Decoy Database Generation

To facilitate the target-decoy FDR strategy, a decoy database was generated in R. Each protein sequence from the original target FASTA file was computationally reversed. The headers of these reversed sequences were modified by adding the prefix "D_" to the accession number to clearly distinguish them as decoy entries. This new set of decoy sequences was then written to a separate file named decoy.fasta.

2.3 Mass Spectrometry Data Processing

The experimental mass spectrometry data was processed using the Comet search algorithm, managed through the SearchGUI interface. Two separate search workflows were executed. In the first, the spectral data was searched against only the target database. In the second, the same spectral data was searched against only the decoy database. The results of each search were exported as compressed pepXML files for subsequent analysis.

For both the target and decoy database searches, a consistent set of parameters was used in Comet. The precursor mass tolerance was set to 10.0 parts per million. The fragment mass tolerance was set to 0.02 Daltons. The digestion enzyme was specified as trypsin with specific cleavage rules, allowing for a maximum of two missed cleavages. Fixed modifications included carbamidomethylation of cysteine. Variable modifications included oxidation of methionine and acetylation of the protein N-terminus. For the initial global FDR analysis, the search was configured to return only the top-ranked peptide match for each spectrum.

3 Global False Discovery Rate Analysis

3.1 Score Distributions of Separate Target and Decoy Searches

The results from the separate target and decoy searches were loaded into R. The primary score used for evaluation was the Comet cross-correlation score, or XCorr. To visualize the overall quality of the matches from each search, probability density plots of the XCorr scores were generated for all rank-1 PSMs.

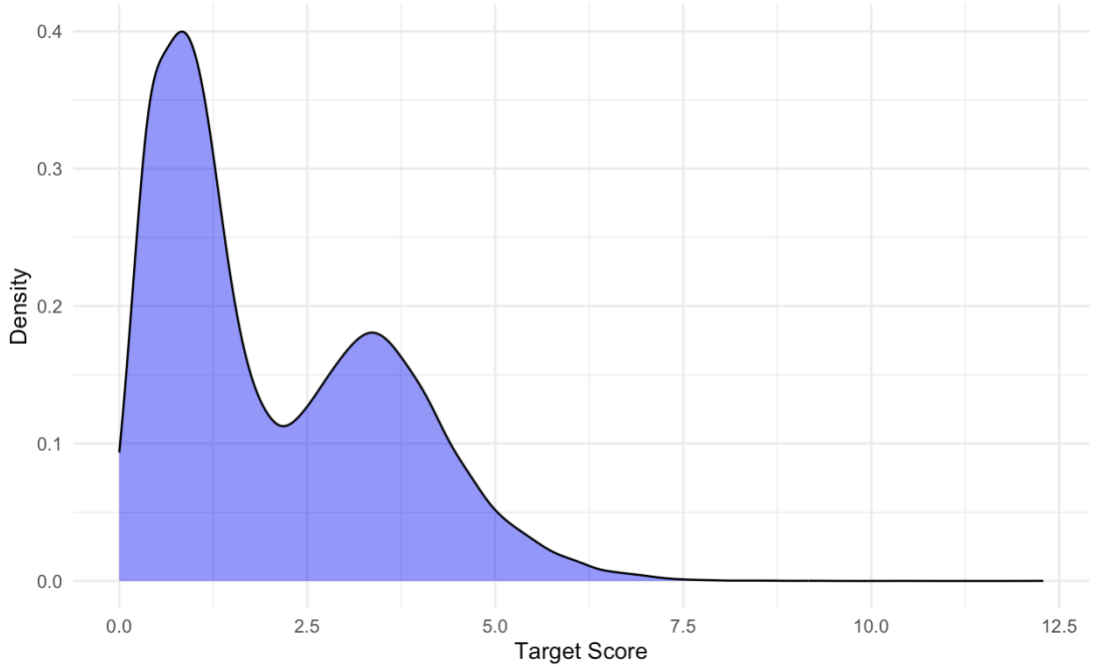

Figure 1: Probability density of Comet XCorr scores for rank-1 PSMs from the target-only database search.

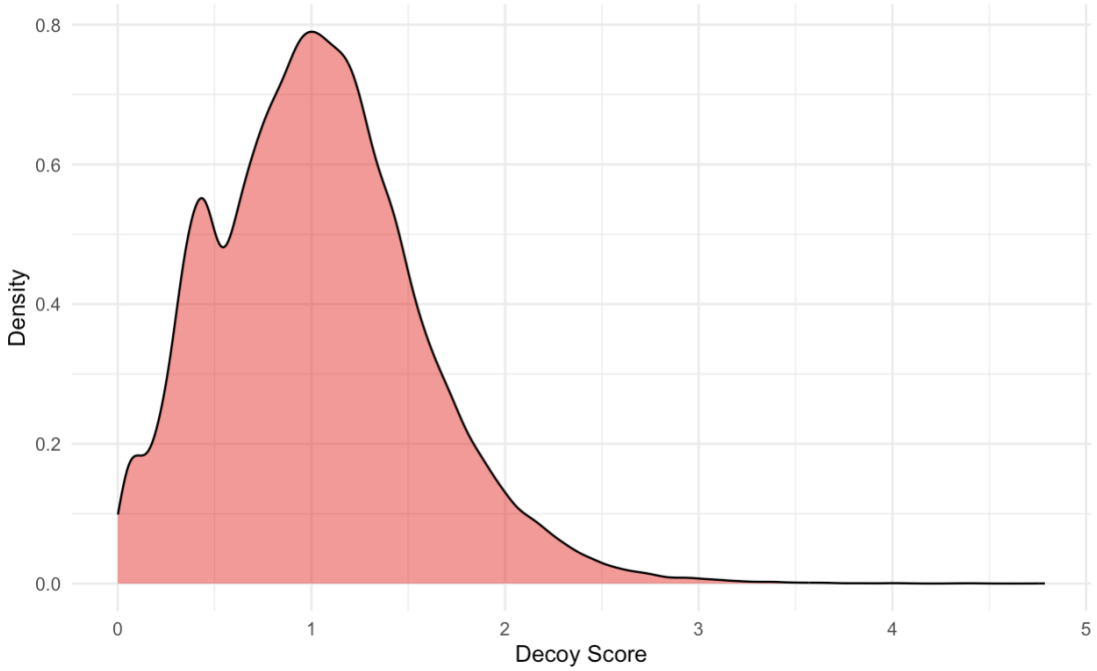

Figure 2: Probability density of Comet XCorr scores for rank-1 PSMs from the decoy-only database search.

From Figures 1 and 2 we can see that the score distribution of the two searches is different. In the target-only database search, the scores follow a bimodal distribution with peaks around values of 1 and 3. A large portion of the scores are higher than 2.5.

On the other hand, in the decoy-only database search, the scores follow a uni-modal distribution centered around the value 1. We can see that almost all the scores have values below 3.

3.2 Emulation of a Concatenated Database Search

To simulate a concatenated database search, the results from the separate target and decoy searches were merged into a single matrix based on their unique spectrum ID. For each spectrum, a "rowmax operator", implemented using the pmax function in R, was applied to select the single highest XCorr score, regardless of whether it originated from the target or decoy search. A new column was created to label each winning PSM as either "Target" or "Decoy". This process creates a dataset where only the best match for each spectrum is retained.

Also from Figure 3 it is possible to notice how the scores from the emulated concatenated-database search follow two different distributions.

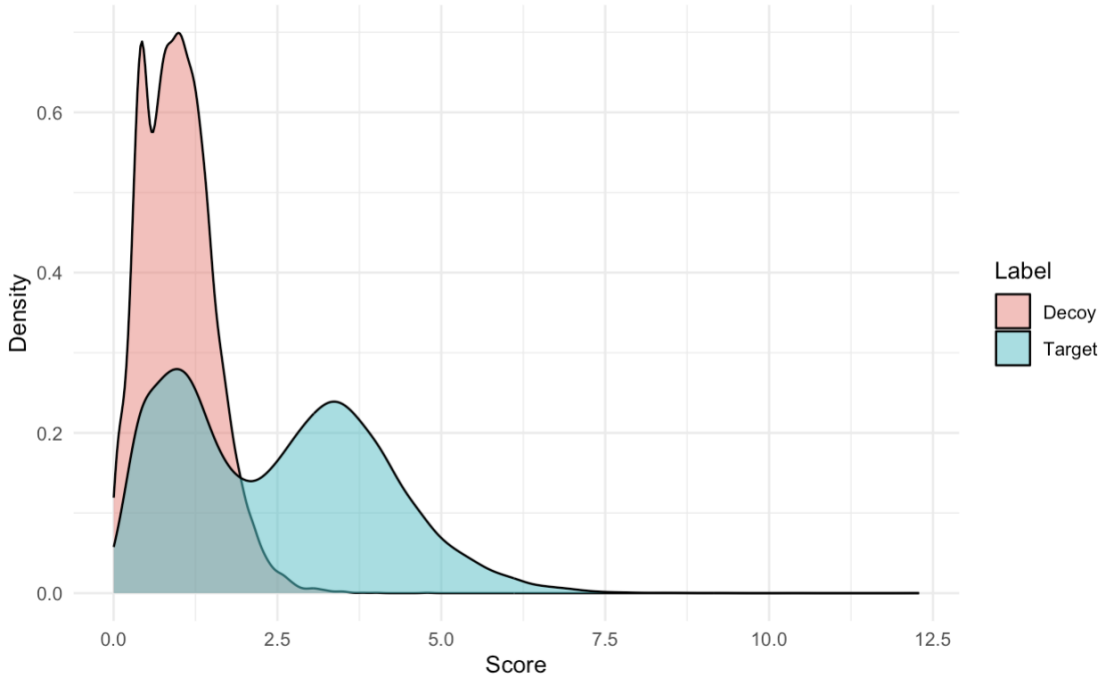

Figure 3: Overlaid probability densities of the winning scores for Target and Decoy PSMs after emulating a concatenated search.

3.3 FDR Calculation for the Emulated Concatenated Strategy

The False Discovery Rate was calculated for the emulated concatenated dataset. The list of PSMs was first sorted in descending order based on the final score. The list was then traversed from top to bottom, and at each position, the cumulative counts of target and decoy hits were calculated. The FDR at each position was computed as the ratio of cumulative decoy hits to cumulative target hits. A threshold of 1 percent was applied to determine the number of confident PSMs.

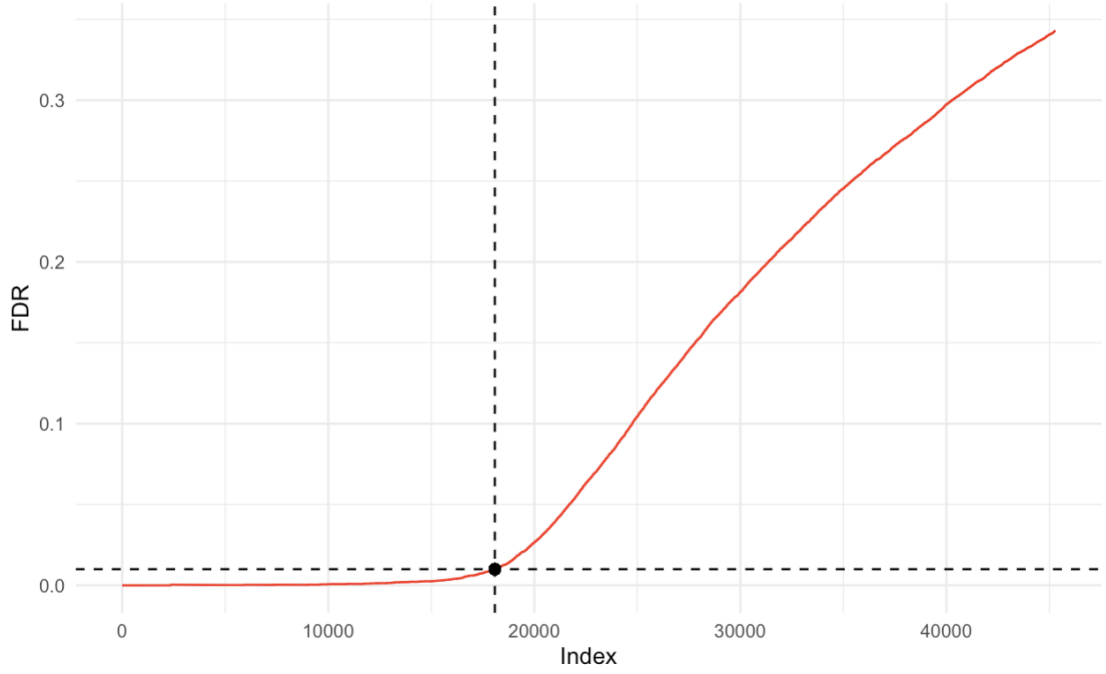

Figure 4: Plot of the calculated FDR as a function of the PSM index in the score-ranked list for the emulated concatenated strategy. The dashed lines indicate the cutoff for a 1 percent FDR.

From Figure 4 we can see that a 1% FDR rate (to be exact, 0.99 %) was reached after selecting the first 18085 peptides. This means we are expected to find only 179 false positive peptide matches out of the first 18085 peptides in

the list.

3.4 Comparison of Concatenated and Un-concatenated FDR Strategies

To assess the effect of the search strategy, an un-concatenated FDR calculation was also performed. In this approach, the full lists of rank-1 target and decoy PSMs were combined without selecting the winning match. This combined list was then sorted by score, and the FDR was calculated in the same manner as for the concatenated strategy. The final number of confident PSMs identified at a 1 percent FDR was then compared between the two strategies, as displayed in Figure 5. In the un-concatenated database search, a FDR of 1% was reached after the first 14468 peptides. This means that performing an un-concatenated search yields a fewer number of confident peptide matches.

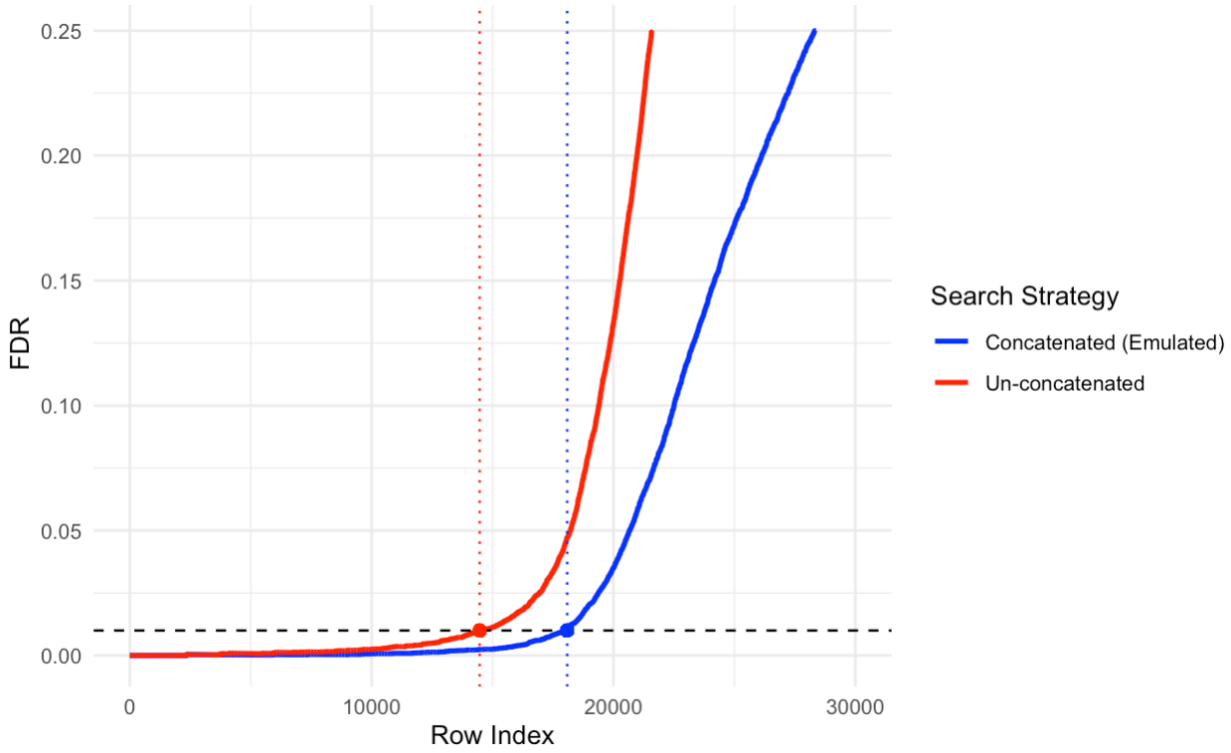

Figure 5: Comparison of FDR curves for the emulated concatenated and un-concatenated search strategies. The x-axis represents the number of accepted PSMs, and the y-axis represents the calculated FDR. The horizontal dashed line indicates the 1 percent FDR threshold. The vertical dotted lines show the total number of confident PSMs identified by each method at this threshold.

3.5 Analysis of Score Behavior and Dependencies

3.5.1 Frequency of Target versus Decoy Wins

In the emulated concatenated analysis, the proportion of spectra for which the target score was greater than or equal to the decoy score was calculated and it was equal to 74.46%. This provides a measure of how often real sequences outperform decoy sequences in a competitive search.

3.5.2 Distribution of Score Differences

The absolute difference between the target and decoy XCorr scores was calculated for every spectrum. A histogram of these differences was generated to visualize the magnitude of separation between the best target and best decoy match.

From Figure 6 is possible to observe how for many PSMs, the difference between the target score and decoy score is not very high, as shown by the big spike around 0.

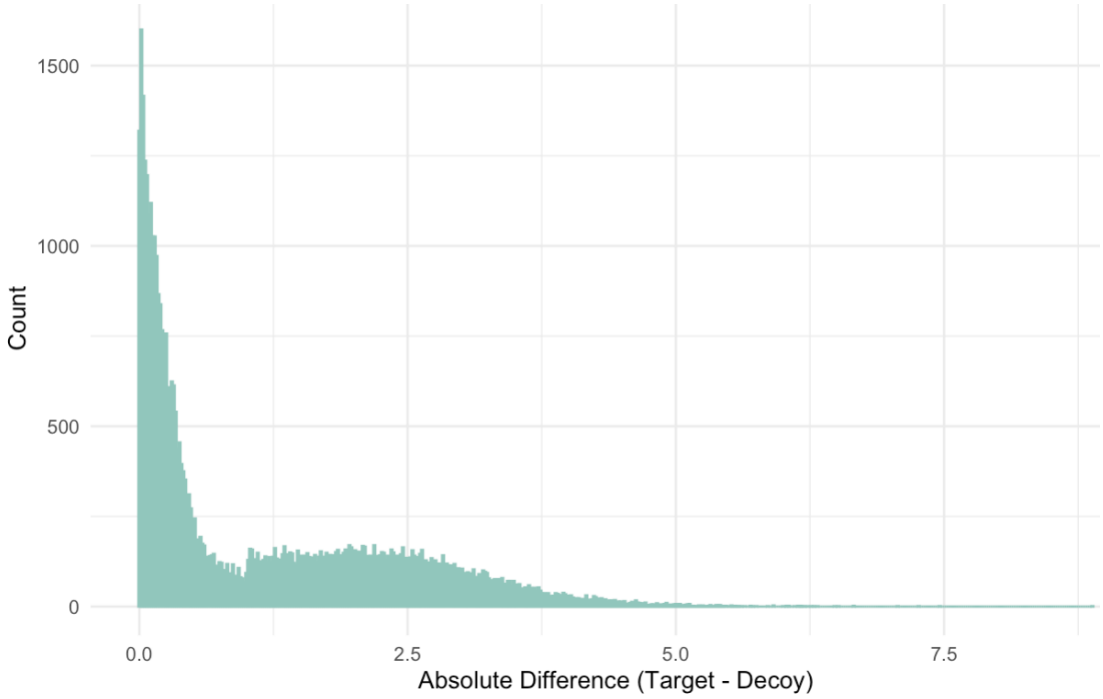

Figure 6: Histogram of the absolute difference between Target and Decoy XCorr scores for each spectrum.

3.5.3 Score Difference as a Function of Rank

To investigate how the competition between target and decoy hits changes with score, the score difference, calculated as Target Score minus Decoy Score, was plotted against the index of the score-sorted PSM list. Figure 7 illustrates whether high-scoring PSMs are more decisively identified as targets compared to low-scoring PSMs.

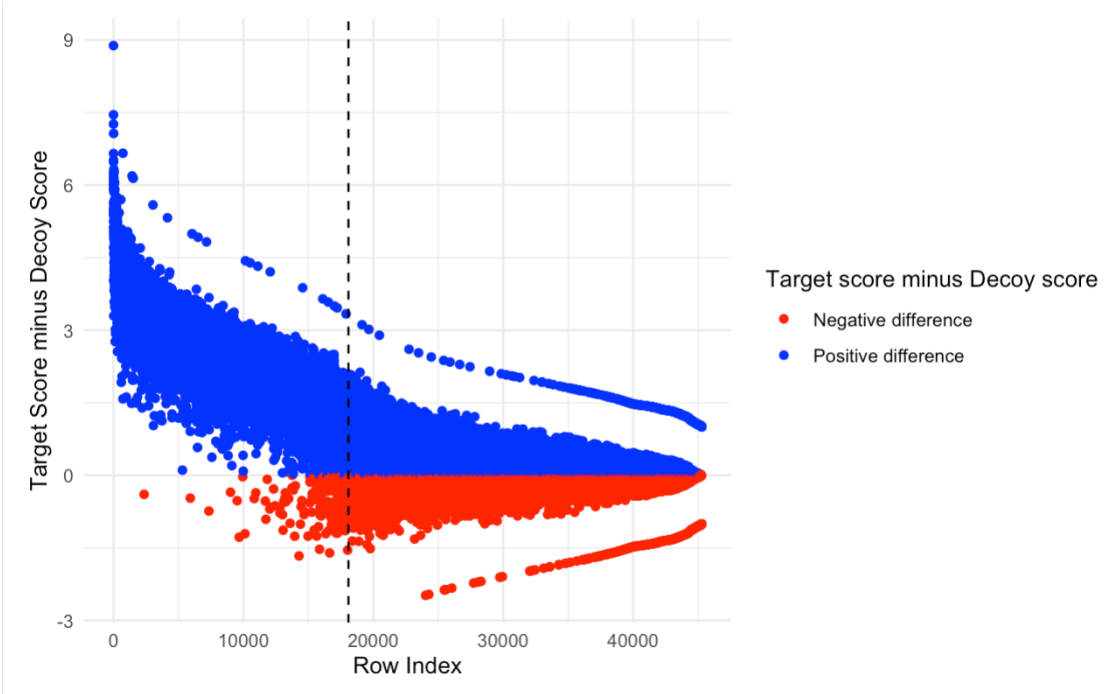

Figure 7: Plot of the score difference (Target minus Decoy) against the rank-ordered PSM index.

For the first 18085 PSM matches, which lie before the vertical dashed line, the target score is consistently higher than the decoy score, with only a minor subset of matches being the opposite. On the other hand, past the dashed line, it is possible to observe almost a mirrored distribution of the difference between the two scores.

4 Local False Discovery Rate Analysis

4.1 Creation of a High-Scoring Spectrum Subset

To investigate per-spectrum score distributions, a subset of the data was created. The top 100 spectra with the highest final scores from the emulated concatenated analysis were identified. Their scan numbers were extracted and used to filter the original mzML file, creating a new, smaller spectral data file containing only these 100 high-quality spectra. A new database search was then performed on this subset file, with Comet parameters adjusted to return the top 100 peptide matches for each spectrum and a wider precursor tolerance of 100 ppm.

4.2 Per-Spectrum Score Distribution Analysis

The results of the new search, containing 10,000 PSMs in total, were loaded into R. For each of the 100 spectra, the distribution of scores for its 100 candidate peptides was visualized. Jitter plots overlaid with boxplots were used to show the score of the rank-1 match in relation to the distribution of scores for the other 99 "noise" matches, as displayed in Figure 8.

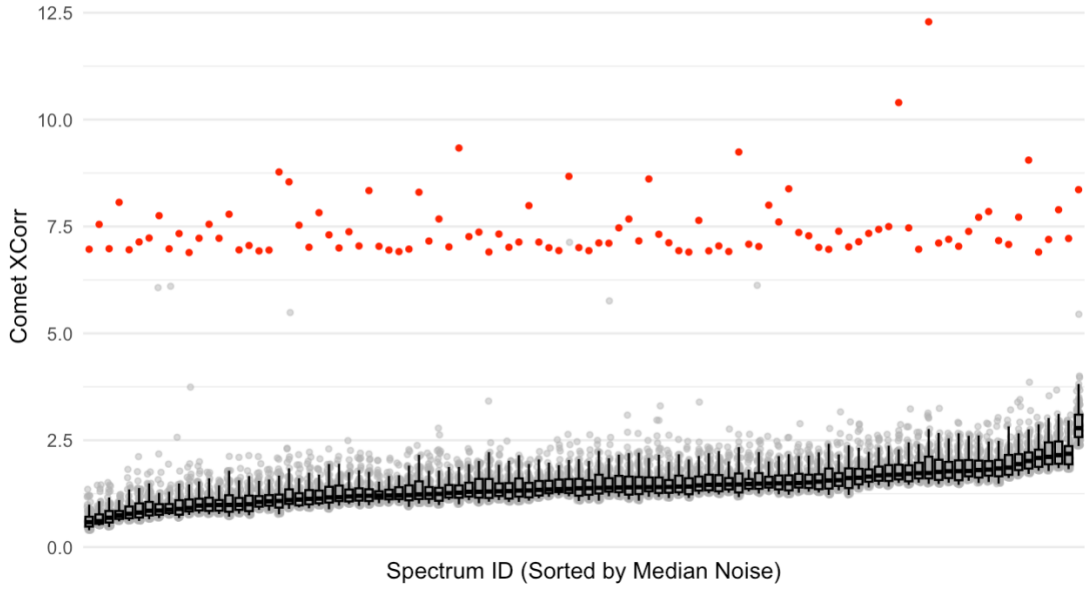

Figure 8: Score distributions for each of the top 100 spectra. Each vertical column of points represents the 100 candidate scores for a single spectrum. The red point indicates the rank-1 match.

This plot beautifully displays how for each of the 100 spectra, the score of the rank-1 match , indicated in red, is a clear outlier in the within-spectrum score distribution. This finding suggests that for each PMS only the rank-1 match should be considered.

4.3 Joint Score Distribution of High-Scoring Spectra

The joint distribution of all 10,000 scores from the local FDR analysis was visualized using a histogram in Figure 9. The distribution of scores for rank-1 matches was colored differently from the distribution of scores for matches ranked 2 through 100. This plot clearly illustrates the separation between the "signal" of the best matches and the "noise" of the incorrect candidates, once again suggesting that only rank-1 matches should be used.

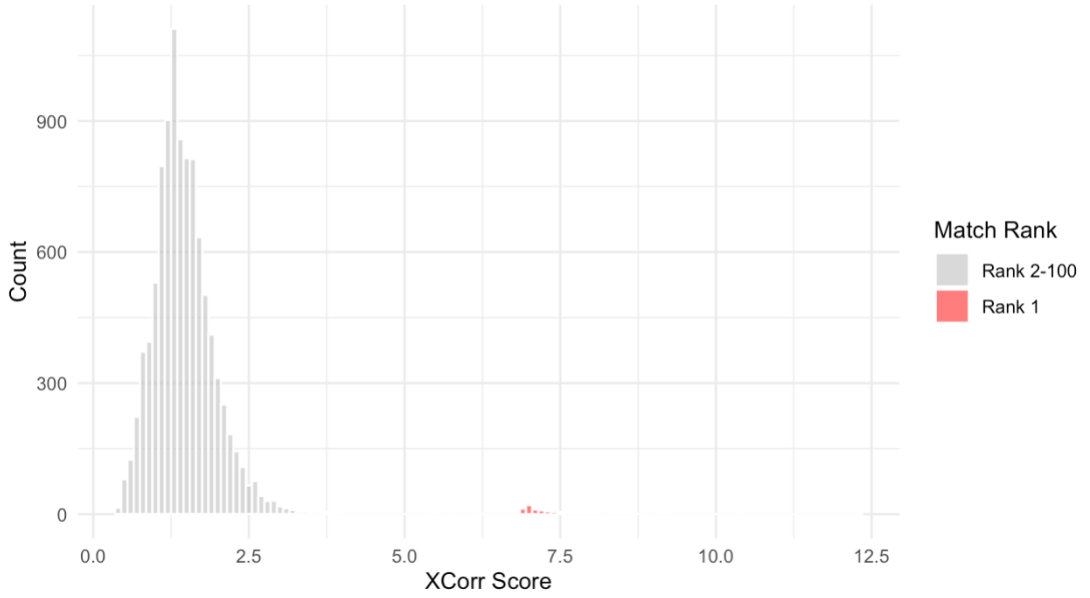

Figure 9: Histogram of all 10,000 scores from the local FDR analysis, showing the distribution of rank-1 scores versus the distribution of scores from ranks 2-100.

5 Discussion

This study aimed to evaluate the impact of different database search strategies on the statistical confidence of peptide-spectrum matches in a shotgun proteomics experiment. The primary finding is that the emulated concatenated search strategy, which incorporates a competition model, identifies a greater number of confident PSMs at a fixed 1 percent False Discovery Rate compared to a more conservative un-concatenated approach, as supported by recent studies [11]. The analysis also highlights the context-dependent nature of search scores, justifying the need for local error rate assessment, a point highlighted in reviews of proteomics error rate estimation [1].

The initial analysis of the separate search results, as shown in Figures 1 and 2, revealed fundamental differences in the score distributions of target and decoy databases. The target score distribution was bimodal, with a large peak centered at a low XCorr score, characteristic of random or incorrect matches, and a second, broader peak at higher scores, representing correct identifications. This suggests a mixture of correct and incorrect matches within the target search space. In contrast, the decoy score distribution was unimodal and centered at a low XCorr score, consistent with the expectation that matches to a decoy database are predominantly random, as described in statistical foundations of FDR in proteomics [12]. Figure 3, which shows the result of the emulated competition, clearly visualizes these two populations, with the decoy hits forming the null distribution and the high-scoring target hits representing the signal.

The central comparison of this report is illustrated in Figure 5, which plots the FDR curves for both the emulated concatenated and the un-concatenated strategies. The un-concatenated approach is demonstrably more conservative, reaching the 1 percent FDR threshold with a smaller number of accepted PSMs. This is because it fails to account for peptide competition, which has been discussed in recent comparisons of search strategies [11]. In this separate-search model, a spectrum can have both a high-scoring target match and a high-scoring decoy match, and both are included in the FDR calculation, causing the rate of false discoveries to increase more rapidly. The emulated concatenated strategy corrects for this by retaining only the single best match for each spectrum, whether it is a target or a decoy. This "winning" match is the only one that contributes to the FDR calculation, which more accurately reflects the process within modern search algorithms and results in a higher number of confident identifications.

Further analysis of score behavior provided insight into the nature of these matches. The histogram of absolute score differences in Figure 6 shows a large number of PSMs where the difference between the best target and best decoy score is close to zero, indicating a high degree of ambiguity for a substantial portion of the dataset. Figure 7 reinforces this by plotting the score difference against the rank-ordered list. For high-ranking PSMs that fall within the 1 percent FDR cutoff, the target score is almost always substantially higher than the corresponding decoy score, indicating decisive identifications. Past the FDR cutoff, the difference becomes more random and symmetrically distributed around zero, confirming that these lower-scoring matches are ambiguous and unreliable.

The investigation into local score distributions highlighted the context-dependent nature of PSM confidence. As shown in Figure 8, for each of the top 100 spectra, the score of the rank-1 match is a clear statistical outlier when compared to the distribution of scores for the other 99 "noise" candidates. This strongly suggests that for high-quality spectra, the best match is decisively better than random chance. However, it is critical to acknowledge that this analysis was performed on a "best-case scenario" dataset, consisting of the 100 PSMs with the highest rank-1 XCorr scores from the entire experiment. While 8 shows a stark and encouraging separation between signal and noise for these spectra, it is reasonable to hypothesize that this separation would become less clear for lower-scoring PSMs. As one moves down the score-ranked list, the distributions for individual spectra would likely become "dirtier", with the rank-1 score being

less of an outlier and sitting closer to, or even within, the distribution of the noise matches. Therefore, while Figure 9 confirms that focusing on the top-ranked peptide is a valid strategy, the clarity of these top 100 results should not be extrapolated to the entire dataset. The ambiguity for lower-scoring PSMs is precisely the problem that formal local FDR and posterior probability calculations aim to solve.

This analysis has several limitations. The study was conducted on a single dataset from a single instrument, and the findings may not be universally applicable to all experimental setups. Further investigation could include stratified FDR calculations based on precursor charge, as suggested in the latest related articles [13], to potentially increase the number of confident PSMs. Additionally, the reasons for decoy wins, whether due to random chance or accidental sequence homology from the reversal process, were not explored through sequence similarity analysis.

In summary, the results support the use of a concatenated database search strategy, or an emulation thereof, as it provides greater statistical power by incorporating a competition model, a conclusion aligned with recent findings in this area [11]. The analysis of per-spectrum score distributions underscores the importance of considering local context when assessing the confidence of an individual PSM, providing a clear rationale for the development and use of local FDR methodologies, which is confirmed by reviews of proteomics error rate estimation [1].

References

[1] A. I. Nesvizhskii. "A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics". In: *Journal of Proteomics* 73.11 (2010), pp. 2092–2123. DOI: 10.1016/j.jprot.2010.08.009.

[2] Yoav Benjamini and Yosef Hochberg. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 57.1 (1995), pp. 289–300. DOI: 10.1111/j.2517-6161.1995.tb02031.x.

[3] M. Choi et al. "ABRF Proteome Informatics Research Group (iPRG) 2015 Study: Detection of Differentially Abundant Proteins in Label-Free Quantitative LC–MS/MS Experiments". In: *Journal of Proteome Research* 16.2 (2016), pp. 945–957. DOI: 10.1021/acs.jproteome.6b00881.

[4] ProteomeXchange Consortium. *ProteomeXchange: Global Coordinated Mass Spectrometry Data Sharing in Proteomics*. European Bioinformatics Institute (EMBL-EBI). URL: https://www.proteomexchange.org/.

[5] Vaudel, Marc, Barsnes, Harald, Berven, Frode S., et al. *SearchGUI: An Open-Source Graphical User Interface for Omics Sets*. CompOmics. URL: https://compomics.github.io/projects/searchgui.

[6] Eng, Jimmy K., Jahan, Thamina, and Hoopmann, Michael R. *Comet: An Open-Source MS/MS Database Search Tool*. University of Washington. URL: https://uwpr.github.io/Comet/.

[7] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

[8] Laurent Gatto, Sebastian Gibb, and Johannes Rainer. *mzR: An R Package for Mass Spectrometry Data*. URL: https://www.bioconductor.org/packages/release/bioc/html/mzR.html.

[9] Hadley Wickham et al. *dplyr: A Grammar of Data Manipulation*. URL: https://dplyr.tidyverse.org.

[10] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. URL: https://ggplot2.tidyverse.org.

[11] Dominik Madej and Henry Lam. "Common Decoy Distributions Simplify False Discovery Rate Estimation in Shotgun Proteomics". In: *Journal of Proteome Research* 24.3 (2025), pp. 1135–1147. DOI: 10.1021/acs.jproteome.1c00600.

[12] Rovshan G. Sadygov, Justin X. Zhu, and Henock M. Deberneh. "Gentle Introduction to the Statistical Foundations of False Discovery Rate in Quantitative Proteomics". In: *Journal of Proteome Research* 24.3 (2025), pp. 1135–1147. DOI: 10.1021/acs.jproteome.7b00170.

[13] J. Ma et al. "Assessment of False Discovery Rate Control in Tandem Mass Spectrometry Analysis Using Entrapment". In: *Nature Methods* 23 (2025), pp. 1907–1914. DOI: 10.1038/s41592-025-02719-x.