PEAKS DB: New Software for Substantially Improved Peptide Identification from Orbitrap ETD Mass Spectrometry

Bioinformatics Solutions Inc.

Jing Zhang[1], Baozhen Shan[1], Lei Xin[1], Bin Ma[2]
[1] Bioinformatics Solutions Inc, Waterloo, ON
[2] University of Waterloo, Waterloo, ON

Novel Aspect

New software significantly improves peptide identification performance on ETD MS/MS data.

Introduction

Two new techniques, Orbitrap and ETD, are being rapidly adopted in mass spectrometry based proteomics. The adoption of these new technologies requires new analytical software to take full advantage of the new data types. In this study we present such new software, PEAKS DB, for peptide identification with Orbitrap ETD MS/MS data. The new software outperforms other tools commonly in use. Moreover, the combination of the new tool with other existing tools together provides even better results.

Methods

The algorithm takes advantage of the high mass accuracy and different fragmentation ions provided by Orbitrap ETD.

Techniques that contribute to the performance improvement include:
(1) A pre-search step to determine the precursor mass error distribution. The recalibrated mass error is used as a feature in the scoring function.
(2) Frequencies of different fragment ion types, including the hydrogen rearrangement ions, are statistically learned and used in the scoring function.
(3) The similarity between the de novo sequencing result and database search result is used as an important feature in the scoring function.
(4) The score is normalized against random peptide matches with the same spectrum. The normalization makes the scores comparable across different spectra.

For better human interpretation of the score, the peptide-spectrum matching score is further converted to a -10lgP score, where P is the P-value. This means the probability that a false identification in the current search achieves the same or better score.

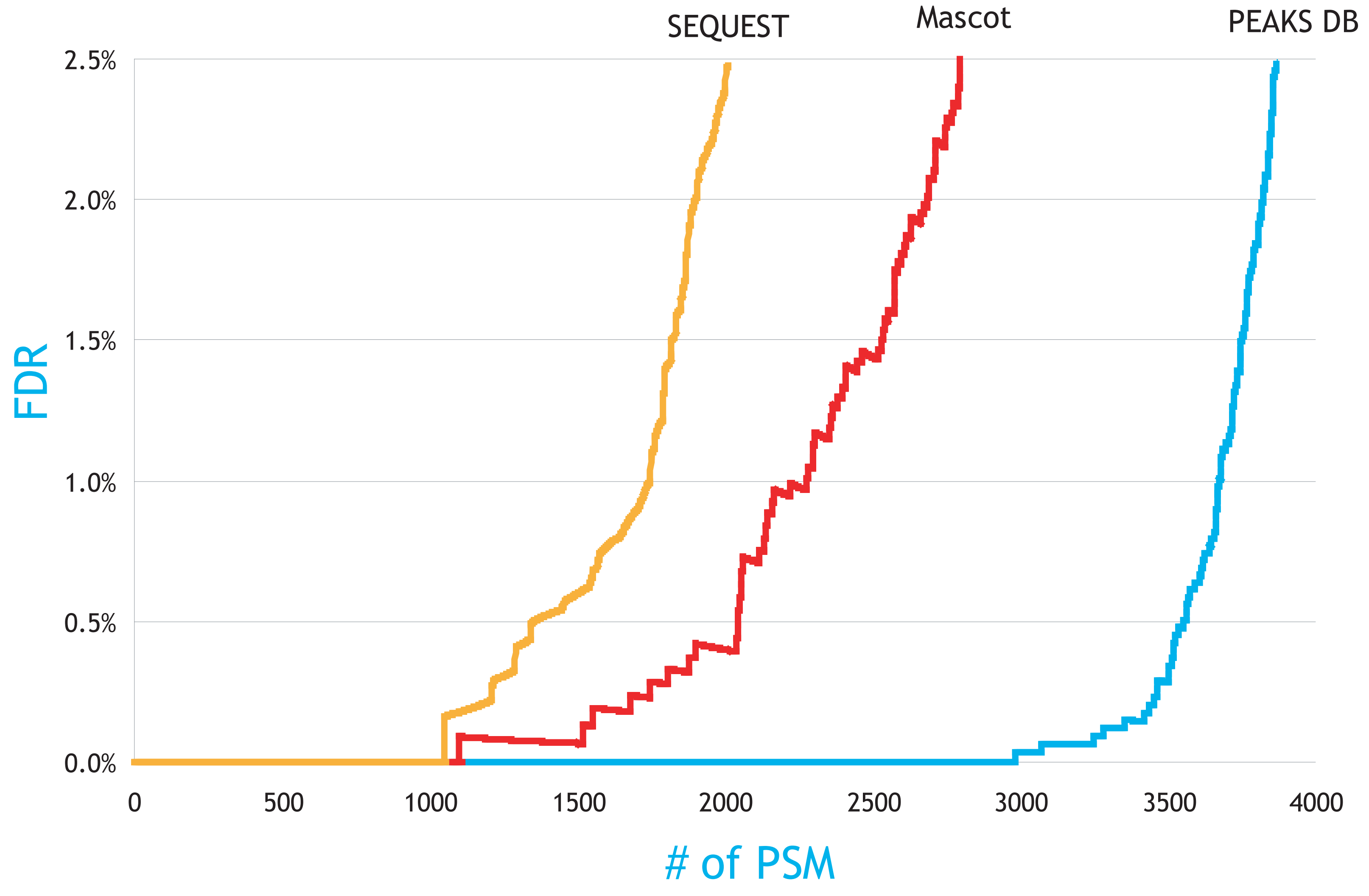
Preliminary Data

The performance of PEAKS DB was studied in two scenarios: (1) when used alone; and (2) when used together with other engines. Two other engines, Mascot and SEQUEST, are examined together with PEAKS DB in this study. The LC-MS/MS test data was collected with a Thermo LTQ-Orbitrap XL ETD instrument on a fraction of Lysine-C digest of yeast lysate. The dataset consists of 8031 MS/MS spectra with ETD and was provided by ABRF in its 2011 study of ETD peptide identification software.

Results

A shuffled decoy was appended to the yeast protein database to determine the false discovery rates (FDR). We use the following PTMs, one fixed PTM: Carboxyamidomethylation of Cys and three variable PTMs: Deamidation of Asn, Oxidation of Met, and Pyro-glu from Q. For SEQUEST, different XCorr score thresholds are used for different charge states, which significantly improved SEQUEST's performance than using the XCorr directly. The number of reported PSM at FDR 1% of each search engine is PEAKS DB (3713) > Mascot (2297) > SEQUEST (1750). The FDR curves of three search engines are given in Figure 1.

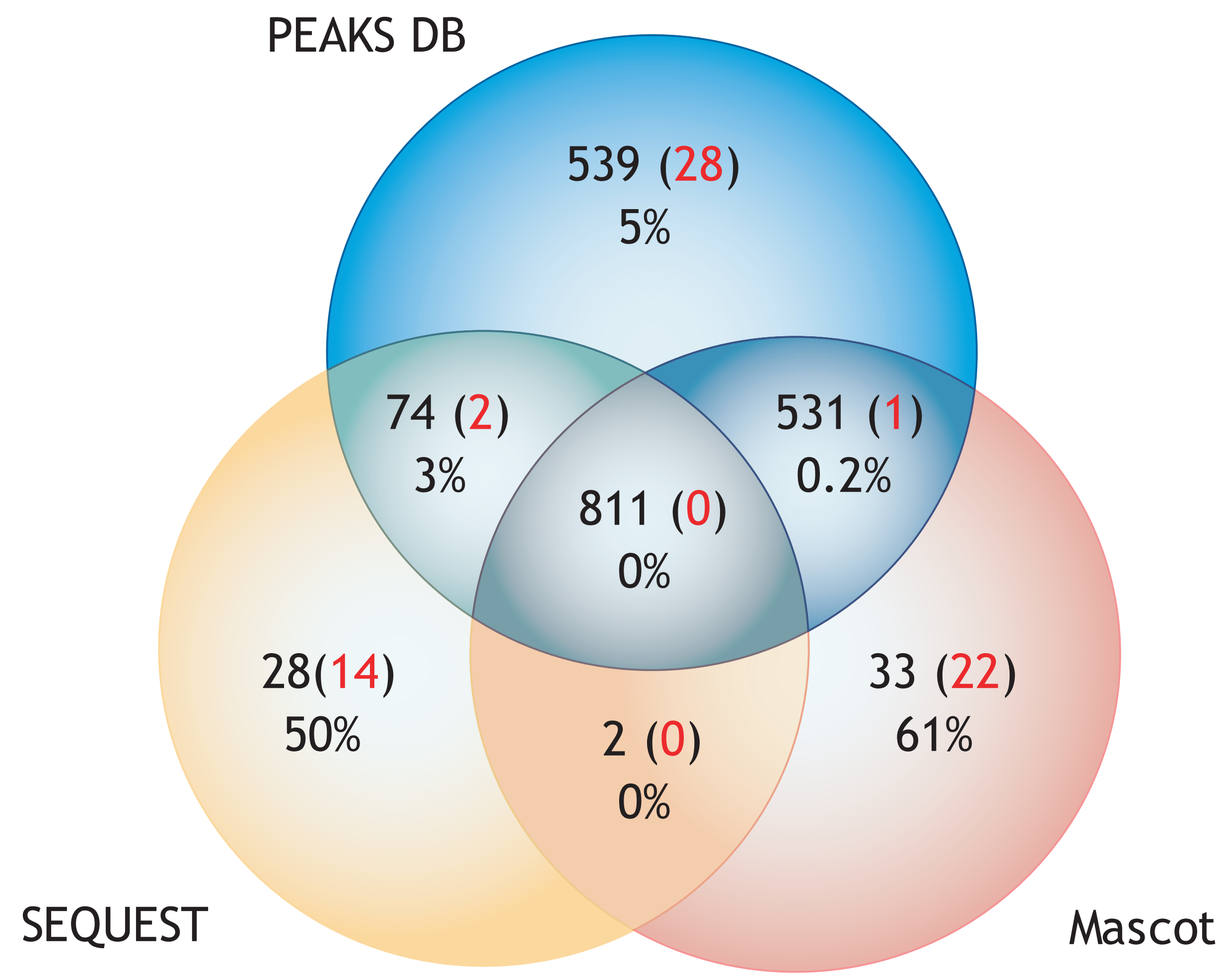
Figure 1: The FDR Curves of Different Search Engines



The performance of using multiple engines together is also evaluated. Results are shown in Figure 2.

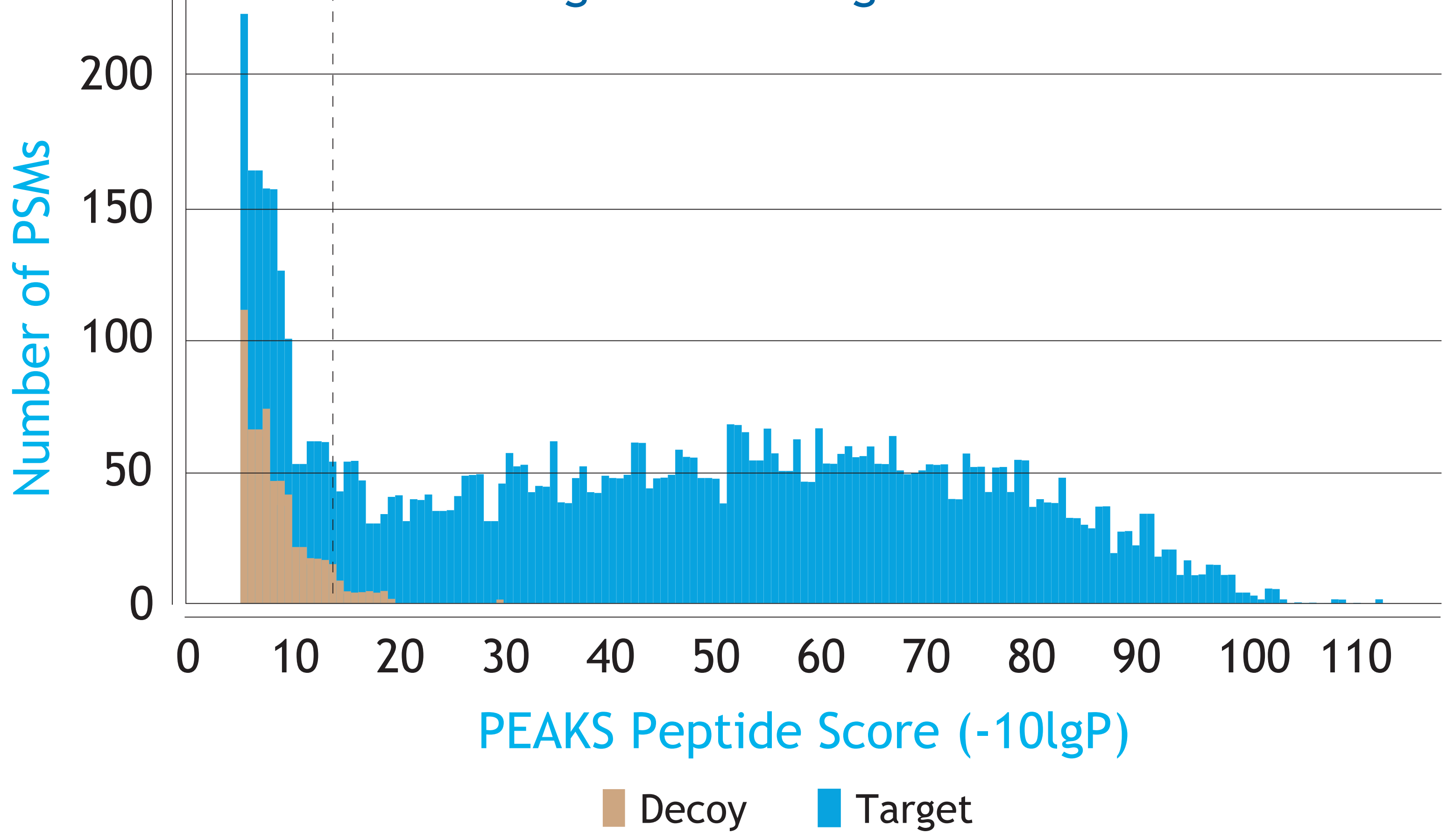
Figure 2: The Venn Diagram of the Number of Distinct Peptides Reported by PEAKS DB, Mascot and SEQUEST

Numbers in red are the decoy hits and the percentiles are the corresponding FDR.



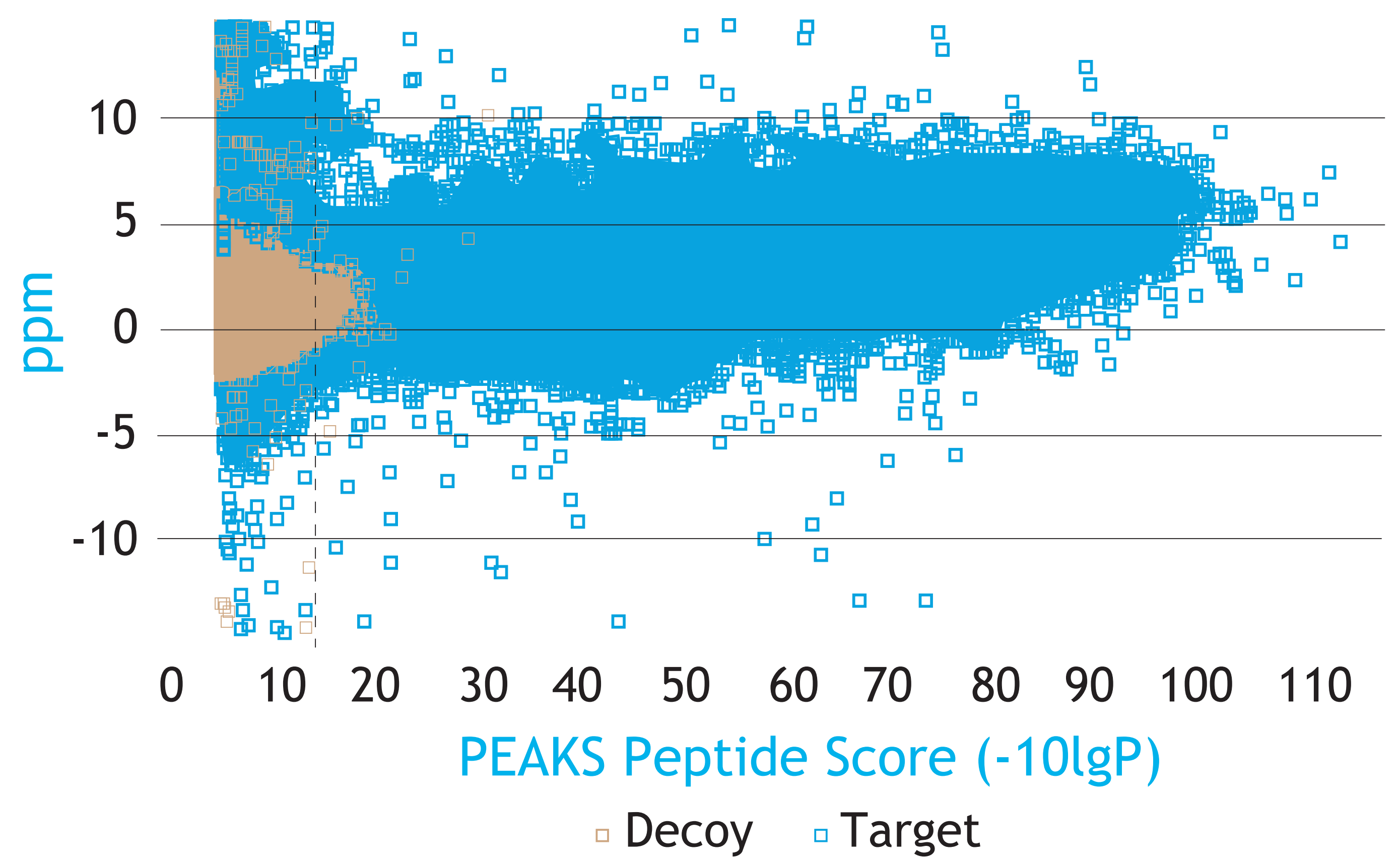
The distribution of PEAKS DB peptide score further shows the ability of to discriminate false hits from true hits. PEAKS DB peptide scores:

Figure 3. Histogram of Score



In Figure 3, the decoy hits are concentrated in the low score region. With the same number of target and decoy hits in the low score region, this indicates the target-decoy FDR estimation is valid.

Figure 4. The Plot of Precursor Mass Error vs. Score



In Figure 4, we can see a clear trend whereby the precursor mass error decreases when the peptide score increases; additionally, the quality of the instrument calibration is visualized.