Supplement to "Tandem mass spectrum identification via cascade search"

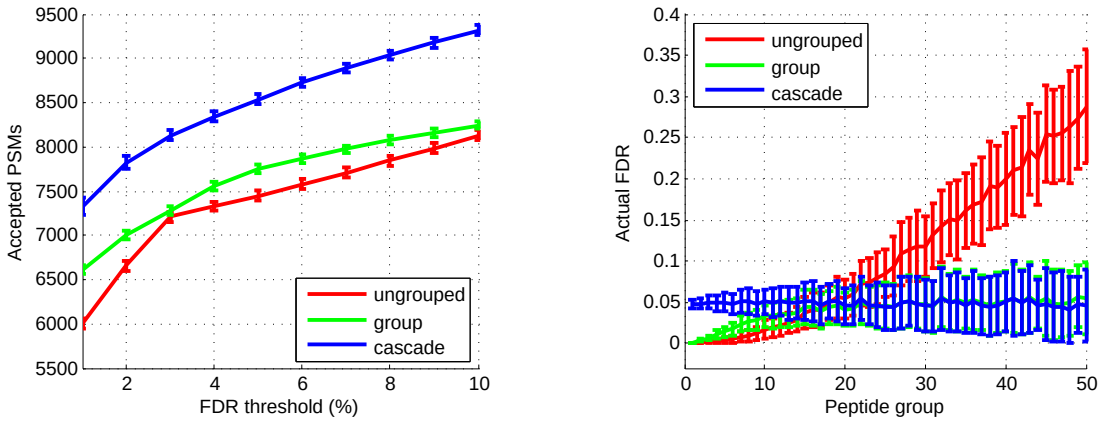
Attila Kertesz-Farkas
Department of Genome Sciences
University of Washington

Uri Keich
School of Mathematics and Statistics
University of Sydney

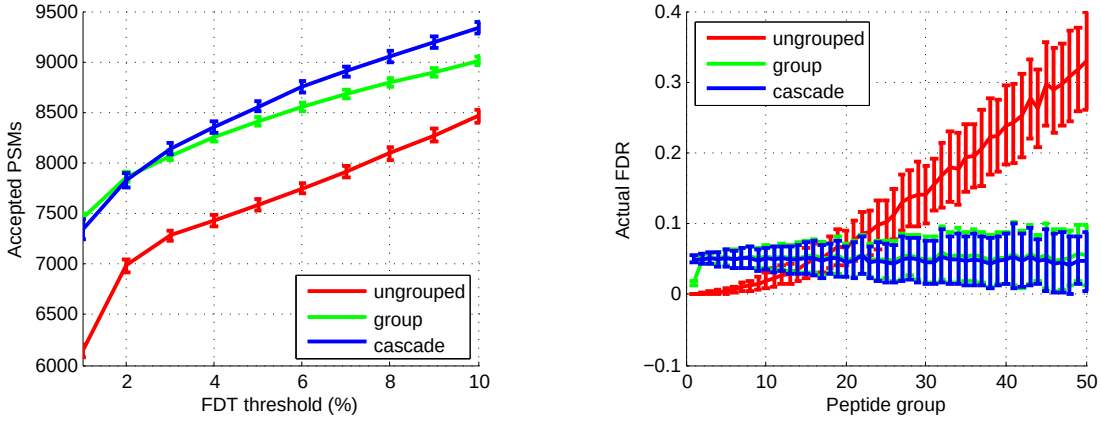
William Stafford Noble*
Department of Genome Sciences
Department of Computer Science and Engineering
University of Washington

February 18, 2015

*Correspondence to wnoble@uw.edu. Phone: 1 206 221 4973

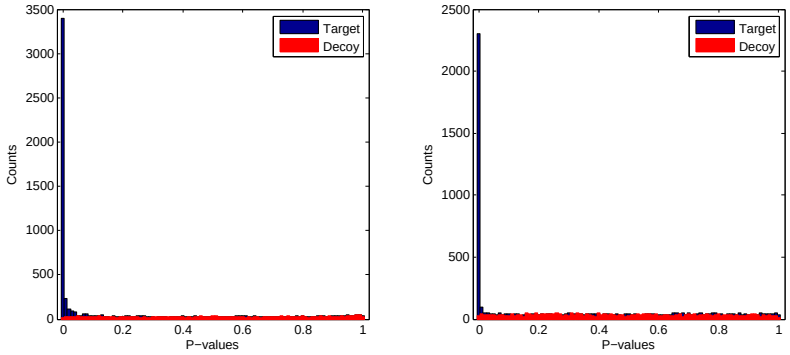


(A) (B)



(C) (D)

Supplementary Figure 1: **Simulation of ungrouped, group and cascade FDR procedures using 50 peptide groups.** 50,000 spectra were searched against 50 peptide groups, where the ith peptide group contained $30i$ candidate peptides, and the number of the identifiable spectra was proportional to $1/i$. This means that the first group contained 2223, the second 1111, etc., while the $50th$ group contained 44 identifiable spectra. All simulations were repeated 100 times, and means and standard deviations are indicated. (A) The figure plots, for each procedure, the number of identified spectra as a function of FDR threshold. (B) The figure plots, for each of the 50 groups, the actual FDR produced by each of the three procedures. (C–D) Similar to (A–B) except that α has been corrected with π_0, as described in the main text.



(A) (B)

Supplementary Figure 2: **Distribution of target and decoy p-values in the Aurum dataset.** (A) A histogram of target (blue) and decoy (red) p-values used in the ungrouped and group FDR methods. Each distribution contains p-values corresponding to matches involving tryptic, semi-tryptic, and non-tryptic peptides. (B) A histogram similar to that in panel (A), but showing p-values from the first iteration of the cascade approach. Target/decoy p-values in subsequent iterations show similar patterns, except that the number of the target p-values close to zero decreases.

Supplementary Table 1: **Number of accepted PSMs at 5% and 10% FDR in the yeast data set.**

	FDR	Tryptic	Semi-tryptic	Non-tryptic	Total
Unrouped		5320	100	449	5869
Group	5%	6812	87	95	6994
Cascade		10861	180	0	11041
Unrouped		5889	124	836	6849
Group	10%	11455	231	14	7629
Cascade		12187	222	0	12409

Supplementary Table 2: **Number of accepted PSMs at 5% and 10% FDR in the Aurum data set.**

	FDR	Tryptic	Oxidized	Methyl	Nt loss	Dioxid	Iodo	Nt acetyl	Total
Ungrouped		2280	573	492	306	163	23	30	3867
Group	5%	2324	588	450	269	156	18	5	3810
Cascade		2450	627	439	268	147	0	0	3931
Ungrouped		2336	606	551	344	193	28	45	4103
Group	10%	2383	619	501	303	171	22	5	4004
Cascade		2572	673	474	297	164	0	0	4180

Supplementary Table 3: **Target-decoy FDR estimates for the Aurum data set**

	Tryptic	Oxidized	Methyl	Nt loss	Dioxid	Iodo	Nt acetyl	Total
Ungrouped	1.05	0.87	5.87	9.00	5.39	26.09	64.52	3.12%
Group	1.55	1.02	2.67	3.70	3.18	10.53	0	1.86%
Cascade	4.40	3.06	2.30	4.85	4.83	10.53	0	4.03%

The FDR was initially estimated at 5% using exact p-values, and then the target/decoy labels were revealed and the FDR was re-estimated for each group. The table reports the target-decoy FDR estimates, as percentages.

Algorithm 1 Controlling FDR using target-decoy analysis. The procedure takes as input a list S of spectra, a corresponding list M of optimal scores, the peptide database D, and the desired confidence threshold α. The procedure returns a list A of Booleans, each indicating whether the corresponding PSM is accepted or not. The procedure generates a decoy peptide set by shuffling (or reversing) each input peptide once. Then this decoy peptide set is used to calculate decoy scores by effectively searching the spectrum set against the union of the decoy and target sets. The subroutine ControlFDRbyEG estimates the FDR using a variant of the target-decoy competition proposed by Elias and Gygi, modified so that it returns only the target PSMs with scores better than the threshold, with the FDR calculation adjusted accordingly.

procedure ControlFDRbyTDC(S, M, D, α)
    $DE \leftarrow$ GenerateDecoy(D)
    $(DM, \_, \_) \leftarrow$ Search(S, DE)
    $M \leftarrow \max(M, DM)$ ▷ A vector of entry-wise maxima.
    $I \leftarrow M < DM$ ▷ I indicates decoys (for simplicity we assume no ties)
    $A \leftarrow$ ControlFDRbyEG(M, I, α) ▷ Control FDR using Elias and Gygi protocol.
    return A
end procedure

Algorithm 2 Controlling FDR using TDC with no peptide groups. The input is a collection S of spectra, a peptide database D, and an FDR threshold α. The subroutine Search(S, D) returns a list E of selected peptides, a list M of scores, where $|M| = |E| = |S|$.

procedure UngroupedFDRbyTDC(S, D, α)
    $(M, \_, E) \leftarrow$ Search(S, D)
    $A \leftarrow$ ControlFDRbyTDC(S, M, D, α)
    return $\{(s_j, e_j, m_j) \mid a_j = 1\}$
end procedure

Algorithm 3 Controlling FDR using TDC with peptide groups. The input is a collection S of spectra, a series $D^1, \ldots, D^n$ of peptide databases, and an FDR threshold α. Note that, unlike Algorithms 2 and 4, this algorithm directly calls ControlFDRbyEG rather than the parent procedure ControlFDRbyTDC.

procedure GroupFDRbyTDC(S, $D^1, \ldots, D^n$, α)
  $\{DE^i \leftarrow \text{GenerateDecoy}(D^i)\}_{i=1}^{n}$;
  $(M, \_, E) \leftarrow \text{Search}(S, D^1 \cup \cdots \cup D^n \cup DE^1 \cup \cdots \cup DE^n)$
  for $i \leftarrow 1 \ldots n$ **do**
    $(S^i, M^i, E^i, I^i) \leftarrow \left\{ \left(s_j, m_j, e_j, e_j \overset{?}{\in} DE^i\right) \mid e_j \in D^i \cup DE^i \right\}$ ▷ I indicates peptide groups
    $A \leftarrow$ControlFDRbyEG(M^i, I^i, α) ▷ Calculate FDR for this group.
    $R^i \leftarrow \{(s_j^i, e_j^i, m_j^i) \mid a_j = 1\}$ ▷ Store return values.
  end for
  return $R^1 \cup \cdots \cup R^n$
end procedure

Algorithm 4 Controlling FDR using TDC with cascaded groups. Like the group FDR algorithm, the input is a collection of spectra, S^0, a series $D^1, \ldots, D^n$ of peptide databases, an FDR threshold α, and a threshold k to abort the procedure when the number of the identification drops below k.

procedure CascadeFDRbyTDC($S^0, D^1, \ldots, D^n, \alpha, k$)
  $R \leftarrow \oslash$
  for $i \leftarrow 1 \ldots n$ **do**
    $(M^i, \_, E^i) \leftarrow \text{Search}(S^{i-1}, D^i)$
    $A^i \leftarrow$ControlFDRbyTDC$(S^{i-1}, M^i, D^i, \alpha)$
    if $|\{i \mid a_j^i = 1\}| < k$ **then**
      break ▷ Abort if the number of identifications is below k.
    end if
    $R \leftarrow R \cup \{(s_j^{i-1}, e_j^i, m_j^i) \mid a_j = 1\}$ ▷ Store return values.
    $S^i \leftarrow \{s_j^{i-1} \mid a_j^i = 0\}$ ▷ Collect unidentified spectra for the next cycle
  end for
  return R
end procedure