# Supplement to "An averaging strategy to reduce variance in target-decoy estimates of false discovery rate"

Uri Keich[1], Kaipo Tamura[2], and William Stafford Noble[2,3]

[1]School of Mathematics and Statistics F07, University of Sydney NSW 2006, Australia, uri@maths.usyd.edu.au, Phone: 61 2 9351 2307
[2]Department of Genome Sciences, University of Washington, Foege Building S220B, 3720 15th Ave NE, Seattle, WA 98195-5065, william-noble@uw.edu, Phone: 1 206 355-5596, Fax: 1 206 685-7301
[3]Department of Computer Science and Engineering, University of Washington

## Supplementary Note 1: Empirical analysis of the power and FDR control of aTDC$_1^+$

Here we report a set of extensive simulations that empirically demonstrate that aTDC$_1^+$ controls the FDR and at the same time, exhibits less variability and delivers at least as many true discoveries as TDC$^+$ does.

### Simulations

For our simulations we adopted our previously published model of dividing the set of spectra into the "native" spectra that were generated by a peptide present in the target database, and the "foreign" spectra that were generated by contaminant peptides, peptide variants that are not in the given database, etc.[1] We simulated three different sizes of spectrum sets: $n = 500$, 10K, and 70K and we let $\pi_1$, the fraction of native spectra, vary in $\{0.1, 0.5, 0.9\}$ ($\pi_0 = 1 - \pi_1$, the fraction of foreign spectra, varied accordingly).

Each of the $n$ spectra had 100 candidates in each database, and each PSM, or a match between the spectrum and one of its candidates, was assigned a label, "true/correct" or "false/incorrect." Matching any candidate peptide against a foreign spectrum obviously yields a "false PSM." Matching a native spectrum against the unique peptide that generated it yields a "true PSM," whereas matching it against any other candidate peptide again yields a "false PSM." We used distinct distributions to model the scores of true and false PSMs.

In our simulations we repeatedly (10K times for each setting of the parameters) and independently drew "target PSM" scores, $\{w(\sigma_i)\}_{i=1}^n$, as well as "decoy PSM" scores, $\{z(\sigma_i)\}_{i=1}^n$. The $z(\sigma_i)$ model the scores of the optimal decoy PSM matches and are drawn according to our null distribution of an optimal PSM score. The same distribution applies to the $n\pi_0$ scores $w(\sigma_i)$ of the optimal target PSMs involving foreign spectra $\sigma_i$. The distribution of $w(\sigma_i)$ for a native $\sigma_i$ is different, and is described in detail below.

We then applied to each of the 10K drawn datasets, simulating the target and decoy PSM scores, both TDC$^+$ and aTDC$_1^+$, where the latter used 3, 10, and 100 competing decoys. We noted the number of discoveries and false discoveries at each nominal FDR level, and we used these to study the accuracy and power of both FDR estimation methods. In particular, for each nominal FDR level of interest $\alpha \in \mathcal{F}$ (the set $\mathcal{F}$ is defined below) we checked if the empirical FDR, which is the mean of the FDP across our 10K samples, essentially coincides with $\alpha$, or whether some consistent bias is observed. Similarly, by looking at the 0.05 and 0.95 quantiles of the FDP at each considered FDR level $\alpha \in \mathcal{F}$ we could gauge the variability in the estimate.

We used the above framework with two different optimal PSM scoring schemes: a calibrated and an uncalibrated one.

### Using calibrated scores

Inverting the order, so that *smaller scores are better*, we use the uniform (0,1) distribution to model the false PSM scores and a beta $B(a = 0.05, b = 10)$ distribution to model the correct PSM scores, $x(\sigma_i)$ for native spectra $\sigma_i$. We denote by $y(\sigma_i)$ the optimal (smallest) score of the best match to the native spectrum $\sigma_i$ among all (99) of its *incorrect* target peptide candidates. Since the distribution of the minimum of $m$ independent uniform (0,1) random variables is a beta $B(a = 1, b = m)$, it follows that $y(\sigma_i)$ has a $B(a = 1, b = 99)$ distribution, and note that $w(\sigma_i) = \min\{x(\sigma_i), y(\sigma_i)\}$.

Similarly, for a foreign spectrum $\sigma_i$, $w(\sigma_i)$ is the minimum of 100 independently drawn uniform (0,1) variates, corresponding to the scores of the 100 matches with its, necessarily false, candidates. As such, it has a beta $B(a = 1, b = 100)$ distribution, and the same applies for the optimal decoy PSM, $z(\sigma_i)$, for *any* spectrum $\sigma_i$.

### Using uncalibrated scores

Our uncalibrated or raw score simulation is based on applying a spectrum specific transformation to our simulated calibrated scores. Specifically, we associated with each spectrum from the yeast data 10K null optimal PSM scores by searching it against that many randomly drawn decoy databases (as described in [2]). More precisely, essentially every spectrum was searched twice, once with a presumed charge of 2 and then with a presumed charge of 3, yielding a total of 68968 sets of 10K null optimal PSM scores.

Using the `R` function `fgev` we separately fitted a generalized extreme value distribution (GEV) to each of these 68968 sets of scores, where the shape parameter was set to 0 so the resulting fit is of a shifted and scaled Gumbel distribution. Applying a Kolmogorov-Smirnov test to gauge the fit between the parametric distribution and the data we kept the 67027 sets of parameters for which the Kolmogorov-Smirnov D statistic was $< 0.05$.

Randomly sampling with replacement we associate with each simulated spectrum one of those pairs of location and scale parameters. Subsequently, for each spectrum $\sigma$, we first draw its target and decoy optimal PSM scores using the calibrated scheme described in the previous section, and then we transform them to raw scores essentially using the quantile function of the Gumbel distribution with the location and scale parameters associated with $\sigma$[1].

## Set of examined FDR values

For computational efficiency our predetermined set of "interesting" FDR values, $\mathcal{F}$, consisted of 120 values defined as: from 0.001 to 0.01 in increments of 0.001, from 0.012 to 0.05 in increments of 0.002 and from 0.055 to 0.5 in increments of 0.005.

# References


[1] U. Keich, A. Kertesz-Farkas, and W. S. Noble. Improved false discovery rate estimation procedure for shotgun proteomics. *Journal of Proteome Research*, 14(8):3148–3161, 2015.

[2] U. Keich and W. S. Noble. On the importance of well calibrated scores for identifying shotgun proteomics spectra. *Journal of Proteome Research*, 14(2):1147–1160, 2015.

[3] U. Keich and W. S. Noble. Progressive calibration and averaging for tandem mass spectrometry statistical confidence estimation: Why settle for a single decoy. In S. Sahinalp, editor, *Proceedings of the International Conference on Research in Computational Biology (RECOMB)*, volume 10229 of *Lecture Notes in Computer Science*, pages 99–116. Springer, 2017.


[1] Strictly speaking, a calibrated score $t$ is replaced with the $1 - t$ quantile rather than the $t$ quantile of the corresponding Gumbel distribution.

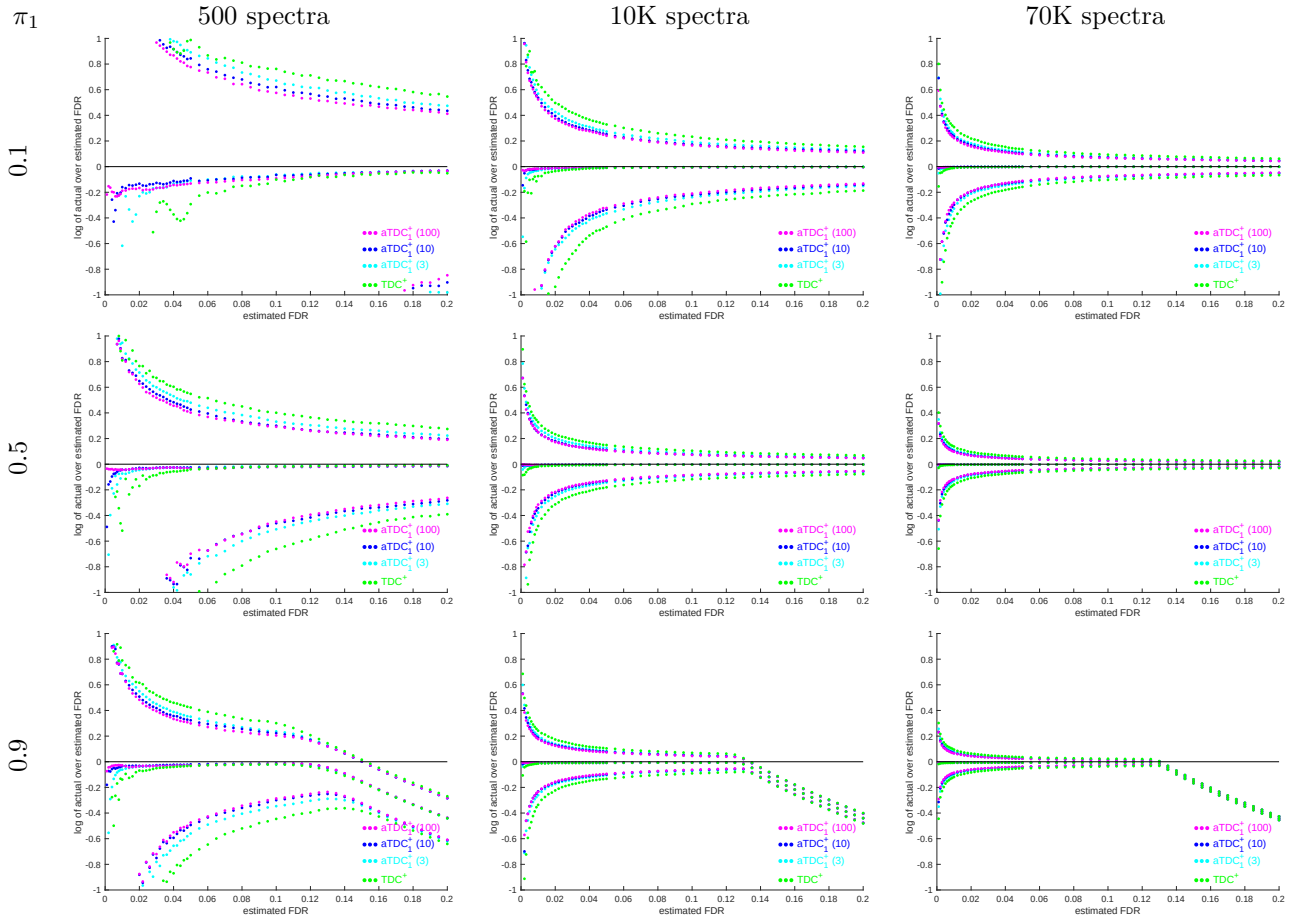


Figure S1: **FDR control: aTDC$_1^+$ and TDC$^+$ (calibrated scores).** Plotted are the (natural) log of the ratios of the mean (empirical FDR, middle curves), 0.05 and 0.95 quantiles (upper and lower curves) of the FDP in the target discovery lists of aTDC$_1^+$ using 3, 10, and 100 competing decoys as well as TDC$^+$ (1 decoy). The FDR is controlled by both methods, with aTDC$_1^+$ exhibiting increasingly reduced variability compared with TDC$^+$. This trend is particularly evident with the smaller spectra sets. Scores are calibrated, and all quantiles are taken with respect to 10K simulation runs using our raw score with 500, 10K, or 70K spectra, 10%, 50%, or 90% of which are native (the $\pi_1$ label to the left of the figure), and 100 candidate peptides per spectrum. The range of the vertical is limited to $[-1, 1]$, so some data points are excluded. The roughly uniform linear fall-off observed in the bottom panels is due to the fact that with 90% native spectra there are only roughly 12-14% false incorrect PSMs (allowing for some native spectra to be incorrectly matched) hence for any FDR threshold larger than that point any method would be conservative. Note that this figure is analogous to Supplementary Figure 3 from[3], except that we used aTDC in the previous version and aTDC$^+$ in the current version.

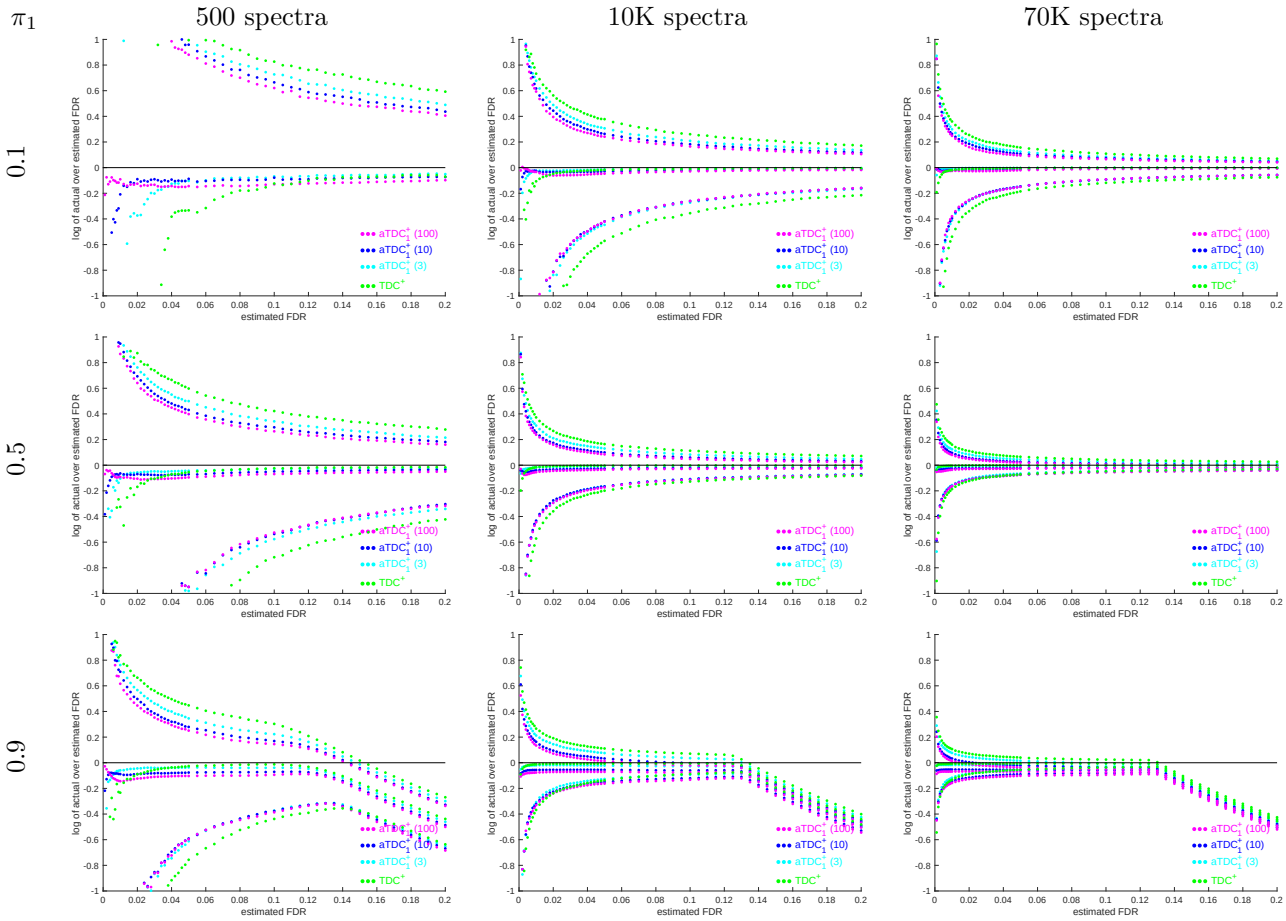


Figure S2: **FDR control: aTDC$_1^+$ and TDC$^+$ (raw scores).** Similar to Supp. Figure S1 but using uncalibrated scores. While aTDC$_1^+$ still exhibits reduced variability compared with TDC$^+$, it also shows a slightly more conservative bias that subtly grows with the number of competing decoys. Still, as seen in Supp. Figure S7, in spite of this conservative bias aTDC$_1^+$ typically reports as many as, or slightly more, *correct* discoveries as does TDC$^+$. Note that the range of the vertical is limited to $[-1, 1]$, so some data points are excluded.

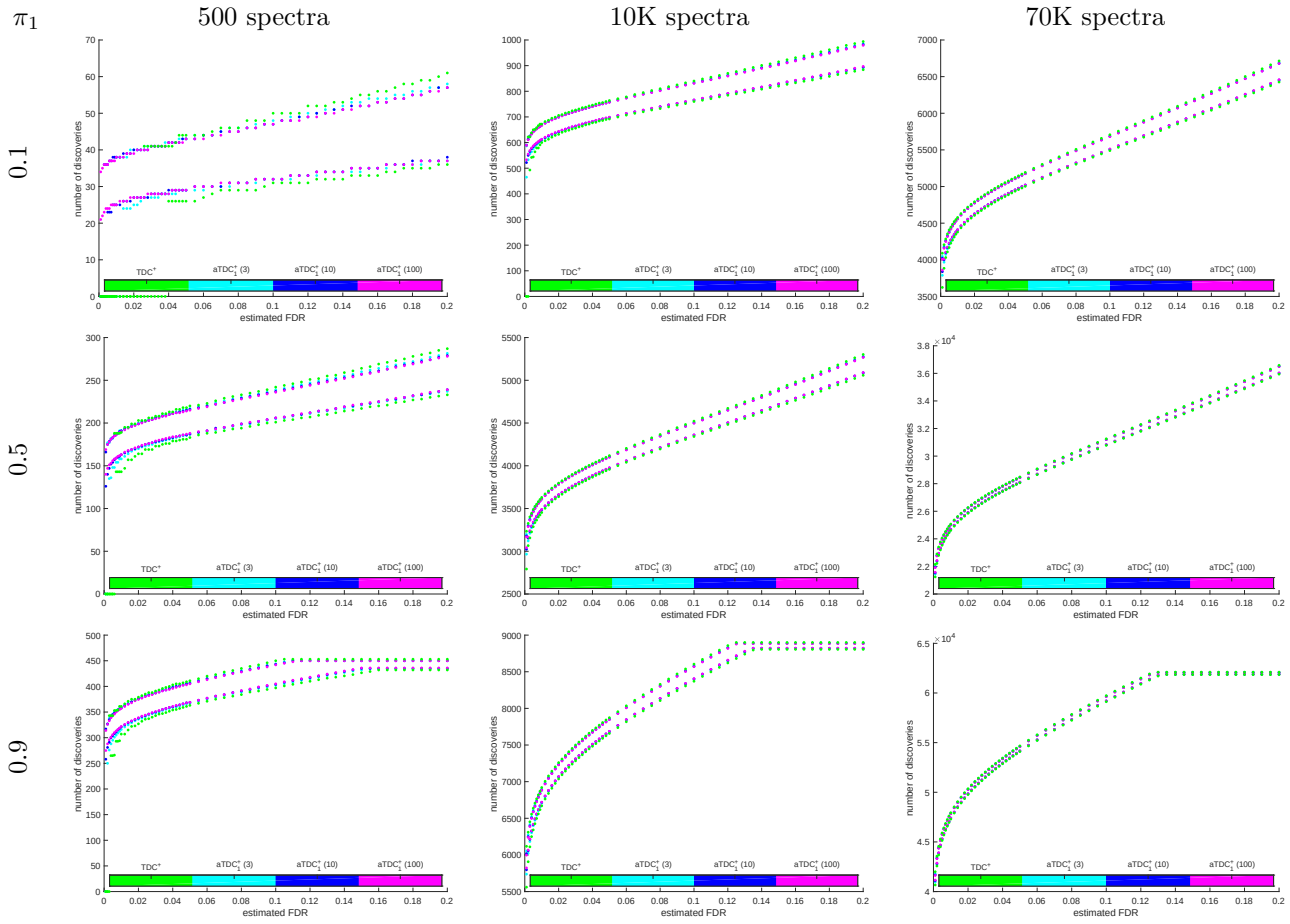


Figure S3: **Variability in number of discoveries: aTDC$_1^+$ vs. TDC$^+$ (calibrated score).** The 0.05 and 0.95 quantiles of the number of target discoveries demonstrate that, consistent with its reduced variability in estimating the FDR, aTDC$_1^+$ exhibits less variability in the number of target discoveries it reports. All quantiles are taken with respect to 10K simulation runs using our calibrated score with 500, 10K, or 70K spectra, 10%, 50%, or 90% of which are native, and 100 candidate peptides per spectrum. Scores are calibrated.

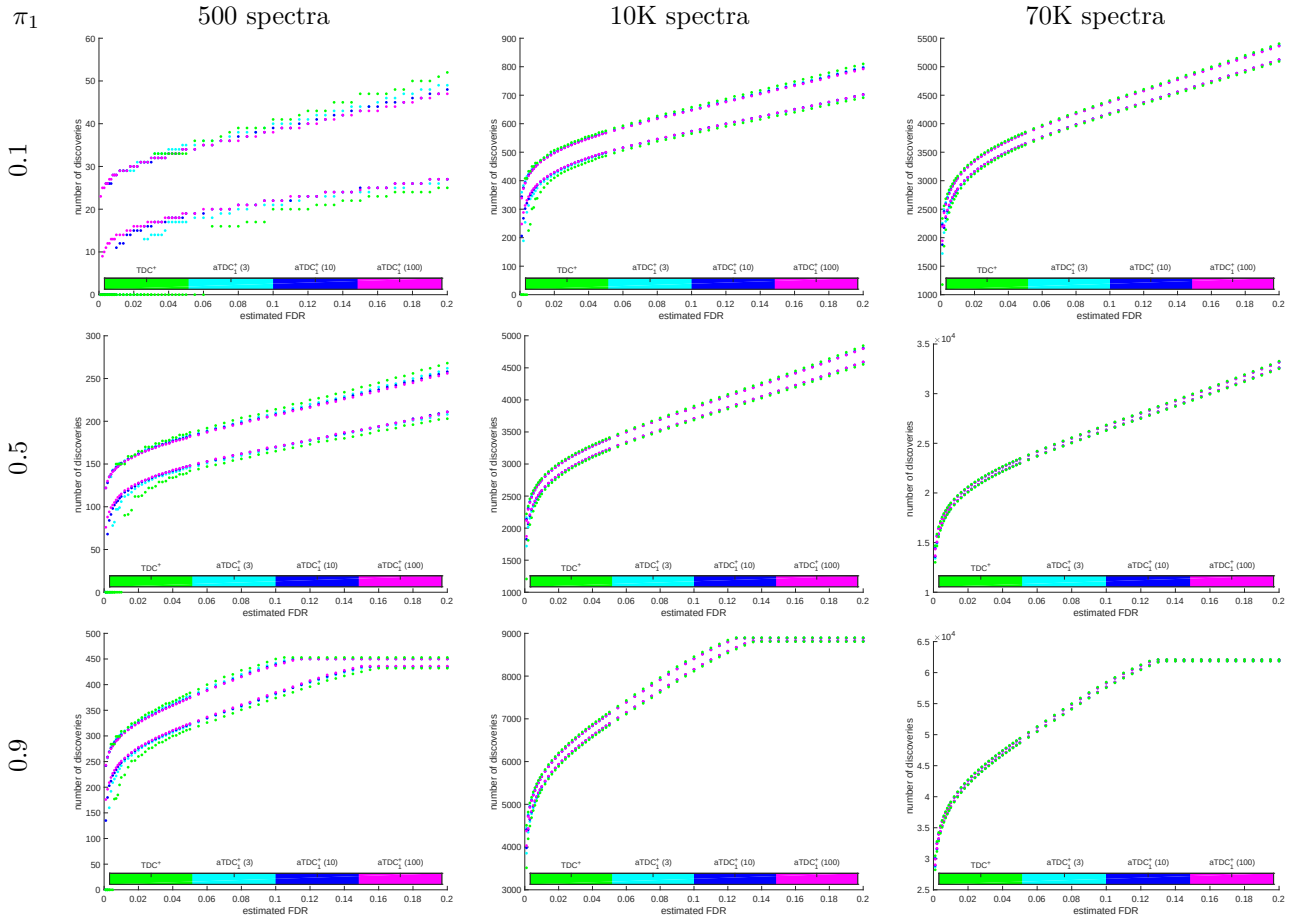


Figure S4: **Variability in number of discoveries: aTDC$_1^+$ vs. TDC$^+$ (raw scores).** Same as Supp. Figure S3 except using uncalibrated scores. We see that aTDC$_1^+$ is able to even further reduce the variability of TDC$^+$. Note that the scales of the $y$-axes vary.

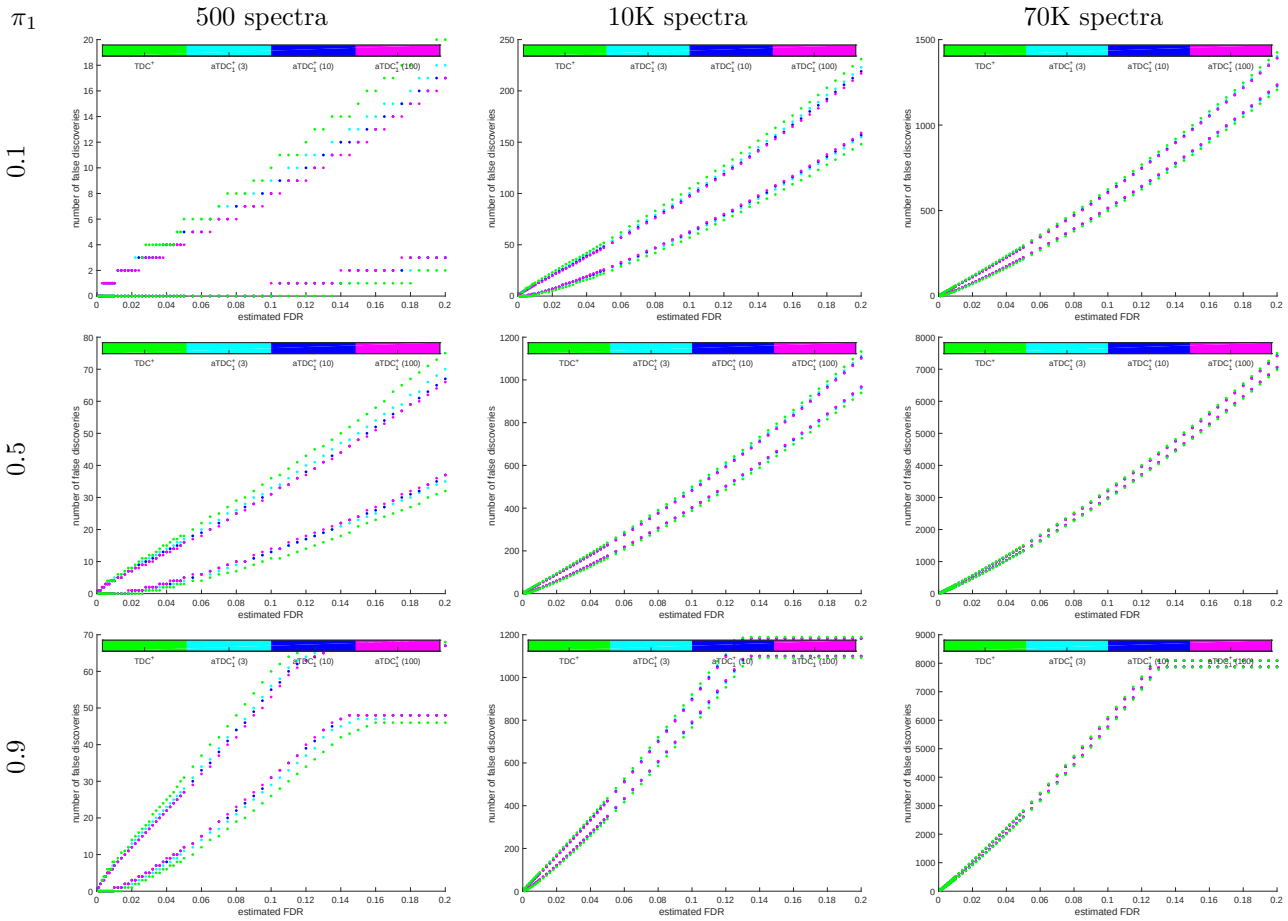


Figure S5: **Variability in number of false discoveries: aTDC$_1^+$ vs. TDC$^+$ (calibrated score).** The 0.05 and 0.95 quantiles of the number of *false* target discoveries offer a slightly different perspective on aTDC$_1^+$'s reduced variability. All quantiles are taken with respect to 10K simulation runs using our raw score with 500, 10K, or 70K spectra, 10%, 50%, or 90% of which are native, and 100 candidate peptides per spectrum. The number of competing decoys was 1 (TDC$^+$), 3, 10, and 100. Calibrated scores.

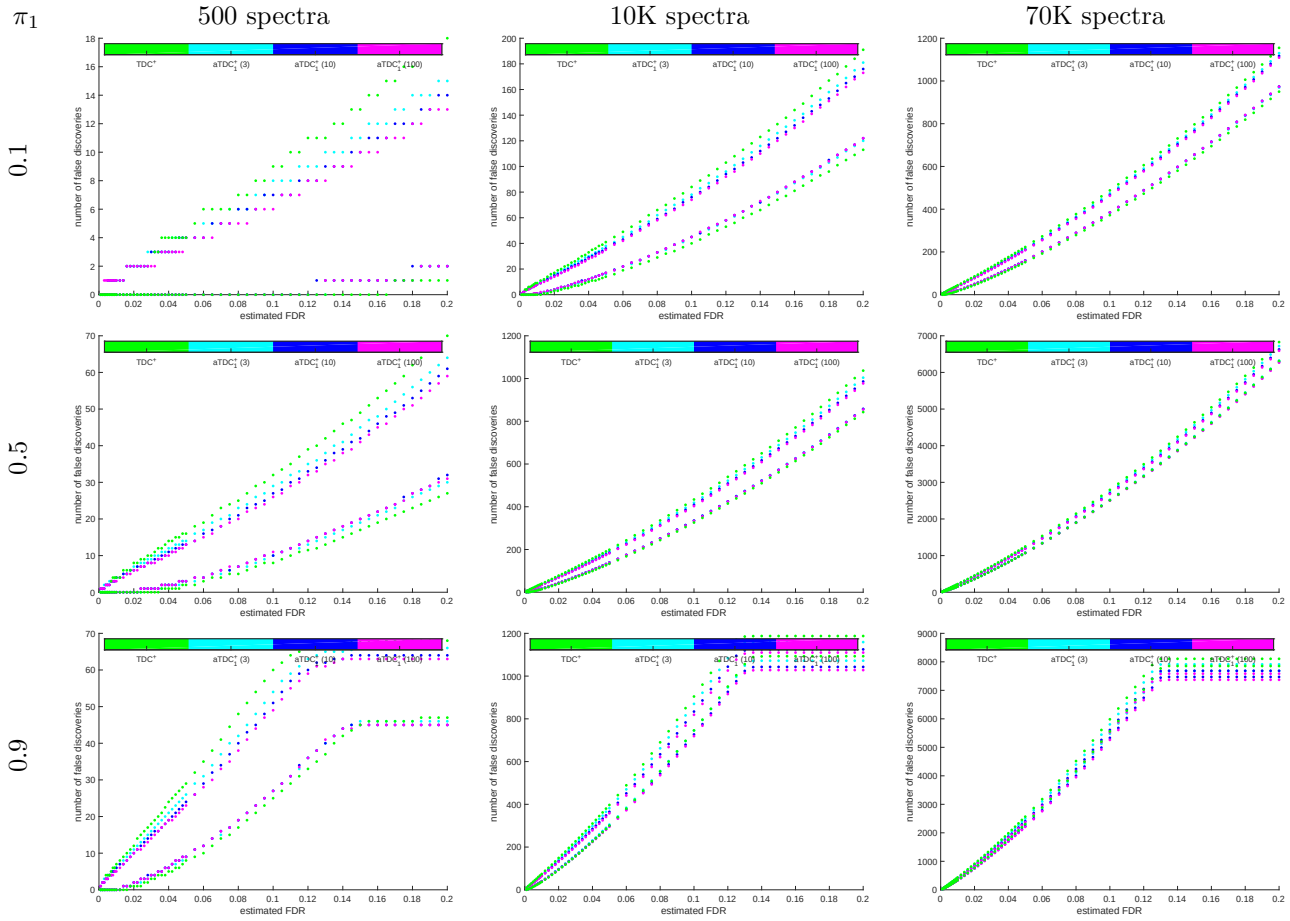


Figure S6: **Variability in number of false discoveries: aTDC$_1^+$ vs. TDC$^+$ (raw scores).** Same as Supp. Figure S5 only with uncalibrated scores. Again we see that aTDC$_1^+$ is able to reduce the variability of TDC$^+$ even further than when using calibrated scores.

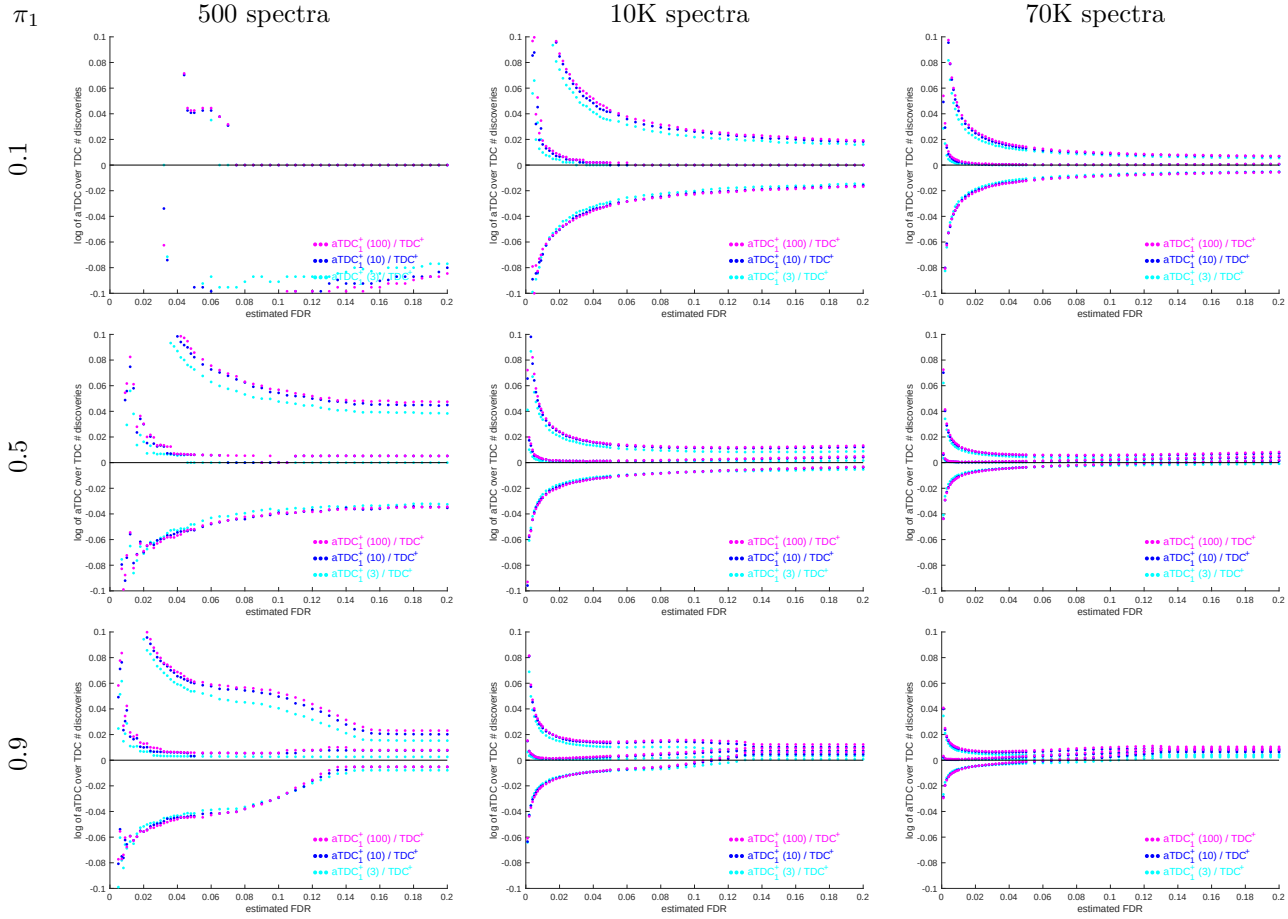


Figure S7: **Power analysis: ratio of number of aTDC$_1^+$ to TDC$^+$ *true* discoveries (raw scores).** The 0.05, 0.5, and 0.95 quantiles of the log of the ratio of the number of aTDC$_1^+$ to TDC$^+$ true discoveries are plotted. Generally, aTDC$_1^+$ offers at least as many true discoveries as aTDC$_1^+$ does with clearly more discoveries for smaller and larger FDR thresholds. This increase is power is more evident with smaller spectra sets as well as sets with a higher proportion of native spectra. Note that these increases in power are specific to non-calibrated scores. When the score is perfectly calibrated, then aTDC$_1^+$ power is comparable to TDC$^+$, though of course it still delivers reduced variability. All quantiles are taken with respect to 10K simulation runs using our raw score with 500, 10K, or 70K spectra, 10%, 50%, or 90% of which are native, and 100 candidate peptides per spectrum. The number of competing decoys was 1 (TDC$^+$), 3, 10, and 100.