# Supplement to "Progressive calibration and averaging for tandem mass spectrometry statistical confidence estimation: Why settle for a single decoy?"

Uri Keich
School of Mathematics and Statistics F07
University of Sydney
uri@maths.usyd.edu.au

William Stafford Noble
Department of Genome Sciences
Department of Computer Science and Engineering
University of Washington
william-noble@uw.edu

December 8, 2017

## 1 Supplementary methods

### 1.1 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.

### 1.2 PC's stoping criterion details

PC's stopping criterion focuses only on the mean increase in the number of discoveries for FDR levels in a range that is specified by the user ($\geq 0.05$ in our experiments). More precisely, let $T(\tau; i)$ denote the number of target discoveries at (TDC / a-TDC estimated) FDR level $\tau$ (Section 2.1), then PC stops the decoy doubling process when the mean of $T(\tau; i)/T(\tau; i-2)$ over our considered list of FDR thresholds $\tau \in \mathcal{F}_0$ is $< \beta$. (Here we used $\beta = 1.005$, corresponding to half a percent mean increase.) The list of FDR thresholds $\mathcal{F}_0$ is defined here as $\mathcal{F} \cap [0.05, 1]$; that is, we restrict our interest to FDR thresholds $\geq 0.05$ (the set $\mathcal{F}$ of FDR thresholds is defined in Supp. Section 1.1).

### 1.3 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

| Variable | Definition |
|---|---|
| $\Sigma$ | set of all spectra (observed) |
| $n$ | number of spectra (observed) |
| $\mathcal{D}^t$ | target database of peptides (user determined) |
| $\mathcal{D}_i^b$ | calibrating decoy database of peptides (user determined) |
| $\mathcal{D}_i^p$ | competing decoy database of peptides (user determined) |
| $n_p$ | number of competing decoy database (user determined) |
| $\mathcal{D}$ | a randomly drawn decoy database of peptides (user determined) |
| $\Sigma_1$ | native spectra (simulations, unobserved) |
| $\Sigma_0$ | foreign spectra (simulations, unobserved) |
| $\pi_1$ | proportion of native spectra (simulations, unobserved) |
| $\pi_0$ | proportion of foreign spectra (simulations, unobserved) |
| $\rho$ | score threshold (user determined) |
| $\alpha, \tau$ | FDR threshold (user determined) |
| $D(\rho)$ | number of decoy discoveries (at level $\rho$) in a search of the concatenated database $\mathcal{D}^t \cup \mathcal{D}$ (observed) |
| $T(\rho)$ | number of target discoveries (at level $\rho$) in a search of the concatenated database $\mathcal{D}^t \cup \mathcal{D}$ (observed) |
| $F(\rho)$ | number of *false* target discoveries (at level $\rho$) in a search of the concatenated database $\mathcal{D}^t \cup \mathcal{D}^p$ (unobserved) |
| $Q(\rho) = F(\rho)/T(\rho)$ | FDP (false discoveries proportion) in a search of the target database (unobserved) |
| $E[Q(\rho)]$ | FDR (false discoveries rate) in a search of the target database (need to control) |
| $\sigma$ | single spectrum (observed) |
| $w(\sigma)$ | score of the best match of $\sigma$ in the target database (observed) |
| $\mathcal{W}$ | set of all optimal target PSM scores (observed) |
| $x(\sigma)$ | score of the match between the native spectrum $\sigma$ and the peptide that generated it (unobserved) |
| $y(\sigma)$ | the score of the match between the native spectrum $\sigma$ and the all other (incorrect) candidate peptides (unobserved) |
| $z_i^b(\sigma)$ | score of the best match of $\sigma$ in the calibrating decoy database $\mathcal{D}_i^b$ (observed) |
| $z_i^p(\sigma)$ | score of the best match of $\sigma$ in the competing decoy database $\mathcal{D}_i^p$ (observed) |
| $\mathcal{Z}^p$ | set of all optimal competing decoy scores (observed) |
| $z(\sigma)$ | score of the best match of $\sigma$ in a randomly drawn decoy (observed) |
| $\mathcal{F}$ | set of 120 considered FDR thresholds from 0.001 to 0.5 (defined in Section 2.1 for computational efficiency) |
| $q_\sigma(s)$ | new primary score for partial calibration (user constructed) |
| $\psi_\sigma(s)$ | new partially calibrated score combining the primary and secondary scores (user constructed) |

Table 1: **V**ariables and their definitions.

**Algorithm 1** a-TDC: return the list of target discoveries and associated estimates of the FDR

- Sort the set of optimal target PSM scores, $\{w(\sigma) : \sigma \in \Sigma\}$ in decreasing order and denote them by $\rho_i$.
- For every competing decoy database $\mathcal{D}_j^p$: apply TDC to the target database $\mathcal{D}^t$ and $\mathcal{D}_j^p$ and note the corresponding number of target, $T_j(\rho_i)$, and decoy, $D_j(\rho_i)$, discoveries at level $\rho_i$.
- Use the above TDC data to compute:
  - $q\left(s; \{z_i^p(\sigma)\}_{j=1}^{n_p}\right)$ for every $s = \rho_i$ and deduce $\psi\left(s; \left\{z_j^p(\sigma)\right\}_{j=1}^{n_p}, \mathcal{W}\right)$
  - $\overline{T(\rho_i)} \coloneqq \sum_{j=1}^{n_p} T_j(\rho_i)/n_p$ and $\overline{D(\rho_i)} \coloneqq \sum_{j=1}^{n_p} D_j(\rho_i)/n_p$
- Initiate: $T = 0$ (cumulative number of target discoveries); Set $\mathcal{T}$, the target discovery indicator, to a logical vector of size $n$ (initiated to all TRUE)
- From the largest target score to the smallest ($i = 1 : n$) do:
  - if $T \leq \overline{T(\rho_i)} - 1/2$ set $T = T + 1$ and skip to next $i$
  - else
    - among all $j \leq i$ for which $\mathcal{T}(j)$ is TRUE choose the one with lowest partially calibrated score $\psi\left(s; \left\{z_j^p(\sigma)\right\}_{j=1}^{n_p}, \mathcal{W}\right)$ — this can be done efficiently at a maximal cost of $O(n_p)$ by keeping a list of the target PSMs with each of the $n_p+1$ possible values of $q\left(s; \{z_i^p(\sigma)\}_{j=1}^{n_p}\right)$.
    - set the $\mathcal{T}$ entry for that chosen target PSM to false and continue to next $i$
- Set the vector $T$ that yields the number of discoveries at level $\rho_i$ to the cumulative sum of the vector $\mathcal{T}$
- Set $\widehat{\mathrm{FDR}}(\rho_i) \coloneqq \overline{D(\rho_i)}/T(\rho_i)$ for all $i$

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 (we also looked at 30 candidates, data not shown), 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 a-TDC, 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 $\alpha \in \mathcal{F}$ 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 also applied partial calibration using up to 2047 calibrating decoys to calibrate both the target and competing decoy scores. The same data was used to virtually apply PC.

We used the above framework with two different optimal PSM scoring schemes: a calibrated and an uncalibrated one, where obviously the partial and progressive calibration procedures were applied only to the data generated using uncalibrated scores.

### 1.3.1 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 $\sigma_i$.

### 1.3.2 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].

### 1.4 Real data analysis

The yeast, worm and *Plasmodium* datasets were processed as described previously [2]. Each spectrum was searched against the corresponding target database, as well as against two sets of independently drawn random (decoy) databases. For the yeast data set, which has low-resolution precursor masses, spectra with multiple charge states were searched with each possible state and the highest scoring match among all possible charge states was selected. The first set of decoy databases was a pool of 10K calibrating databases, and the second constituted the pool of 1K competing databases. Thus, associated with each spectrum was its optimal match against the target database ("target PSM"), as well as 10K calibrating decoy PSMs, and 1K competing decoy PSMs.

We examined the effects of progressively calibrating the scores by iteratively (essentially) doubling the number of calibrating decoys. Starting with 0 calibrating decoys, corresponding to the raw scores, at the $i$th doubling cycle we randomly selected (without replacement) for each spectrum $2^{i-1}$ calibrating decoy scores out of the pool of 10K calibrating PSM scores associated with that spectrum. Thus, after the $i$th doubling for each spectrum we had a total of $2^i - 1$ calibrating decoys scores randomly selected without replacement, which contained the corresponding calibrating decoy set of the previous doubling cycle. Altogether we went through 13 doubling cycles plus an additional last cycle that increased the number of calibrating scores to the maximal possible 10,000. Note that even when using 10,000 decoys, the results here slightly differ from our previously reported results [2] because our previous calibration procedure broke ties randomly rather than using the secondary score.

We next used this spectrum-specific chain of calibrating sets to progressively calibrate the score of the target PSM of the corresponding spectrum as well as the scores of its $n_p$ (here we conentrated on $n_p = 10$) competing decoy PSMs that were randomly drawn from the set of 1K competition decoy scores associated with the same spectrum. Only the first of these $n_p$ (progressively calibrated) competing decoy scores was used when applying TDC, and all $n_p$ were used when applying a-TDC.

This entire process was repeated 2,000 times while independently varying the draws of both calibrating and competing decoys. For each of these 2,000 experiments and for each size of calibrating set in the experiment we noted the number of TDC as well as a-TDC target discoveries at each considered level of FDR $\tau \in \mathcal{F}$.

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

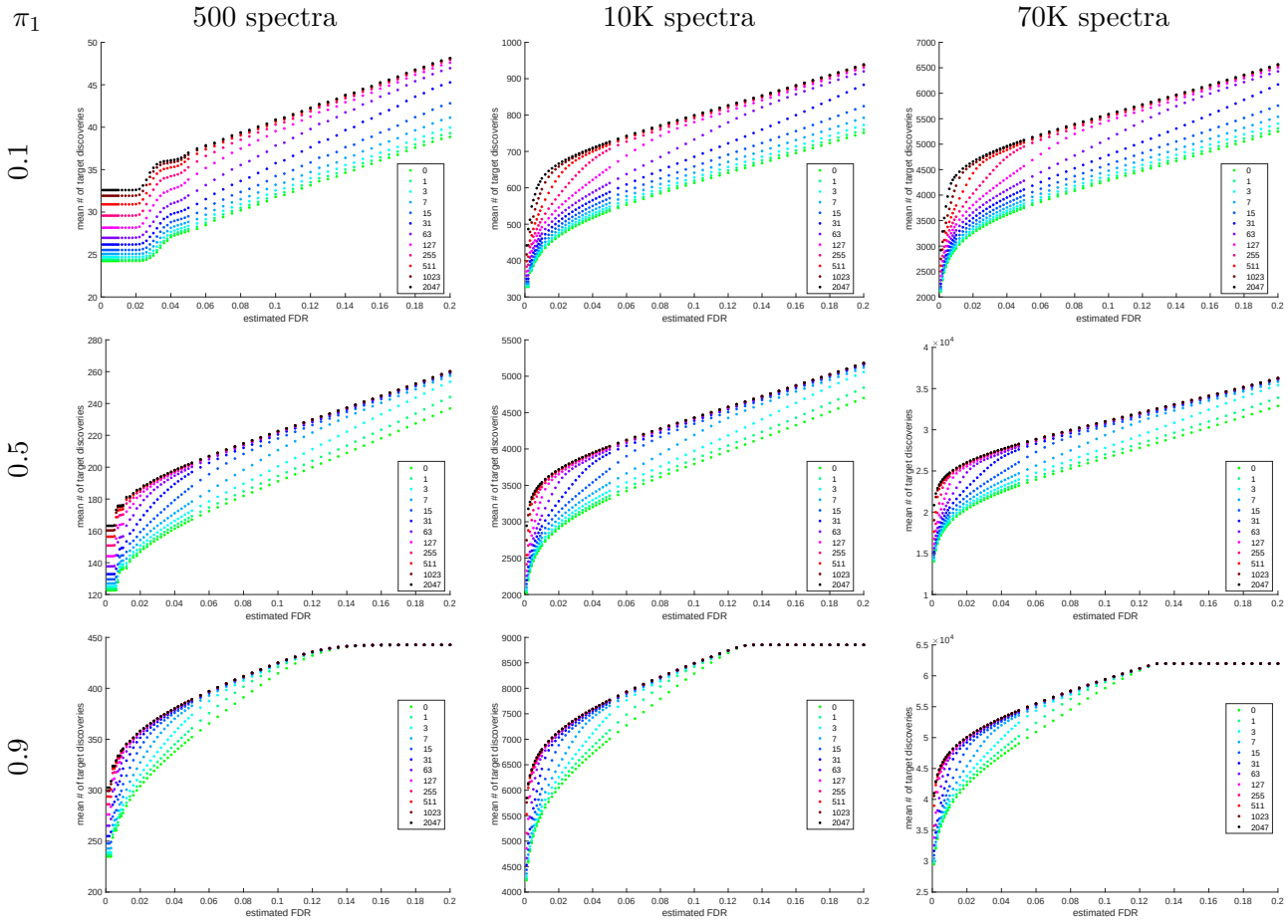


Figure 1: **Partial calibration: Mean number of TDC target discoveries.** The panels show the consistent increase in the mean number of TDC target discoveries as the number of calibrating decoys increases. All means are taken with respect to 10K simulation runs using our raw score with 500, 10K, or 70K spectra, 10%, 50%, or 90% of which are native, and 100 candidate peptides per spectrum. The number of calibrating decoys was varied from 0 to 2047

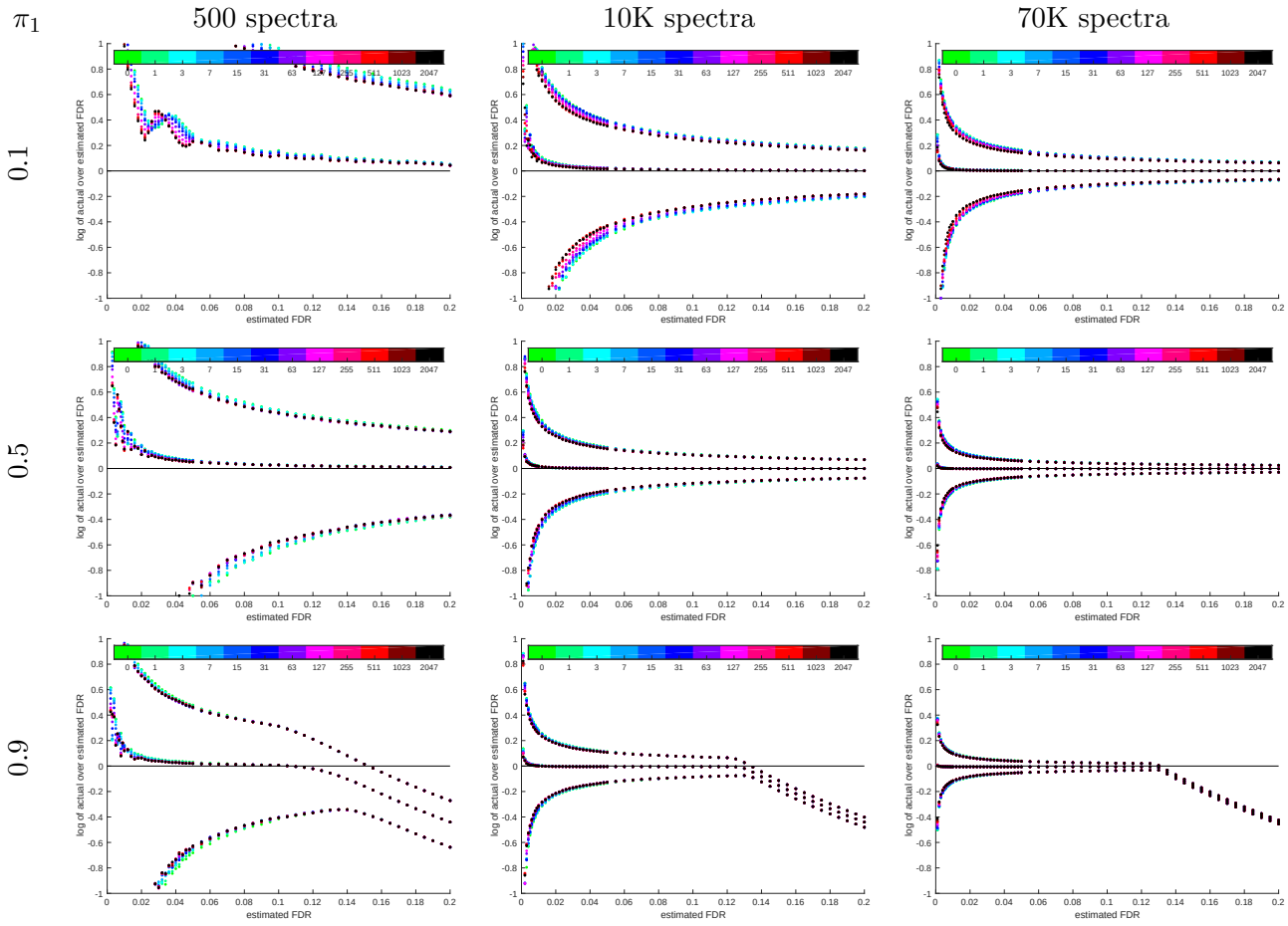


Figure 2: **Partial calibration: FDR control (TDC).** The panels study the accuracy (and variability) of the FDR estimation using TDC. Except for small spectrum sets and small FDR values there is little change in terms of the bias when increasing the number of calibrating decoys: the set of middle curves, which corresponds to the log of the ratio between the empirical FDR (average of the FDP with respect to 10K samples) and the nominal FDR level, essentially coincide for all considered numbers of calibrating decoys. Moreover, TDC seems asymptotically unbiased as the number of spectra increases. Still, for small spectrum sets and/or small FDR values TDC is liberally biased as we pointed out previously [1]. The set of lower and upper curves in each panel correspond to the log of the ratio of the 0.95 and 0.05 quantiles of the FDP to the nominal FDR level across 10K samples. For smaller spectrum sets and particularly for a smaller fraction of native spectra $\pi_1$ we observe a small reduction in the variability of the estimated FDR. Note that for $\pi_1 = 0.9$, the FDP in the entire set is smaller than the limit of the $x$-axis, hence the observed abrupt change to essentially linear conservative bias. 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 calibrating decoys was varied from 0 to 2047.

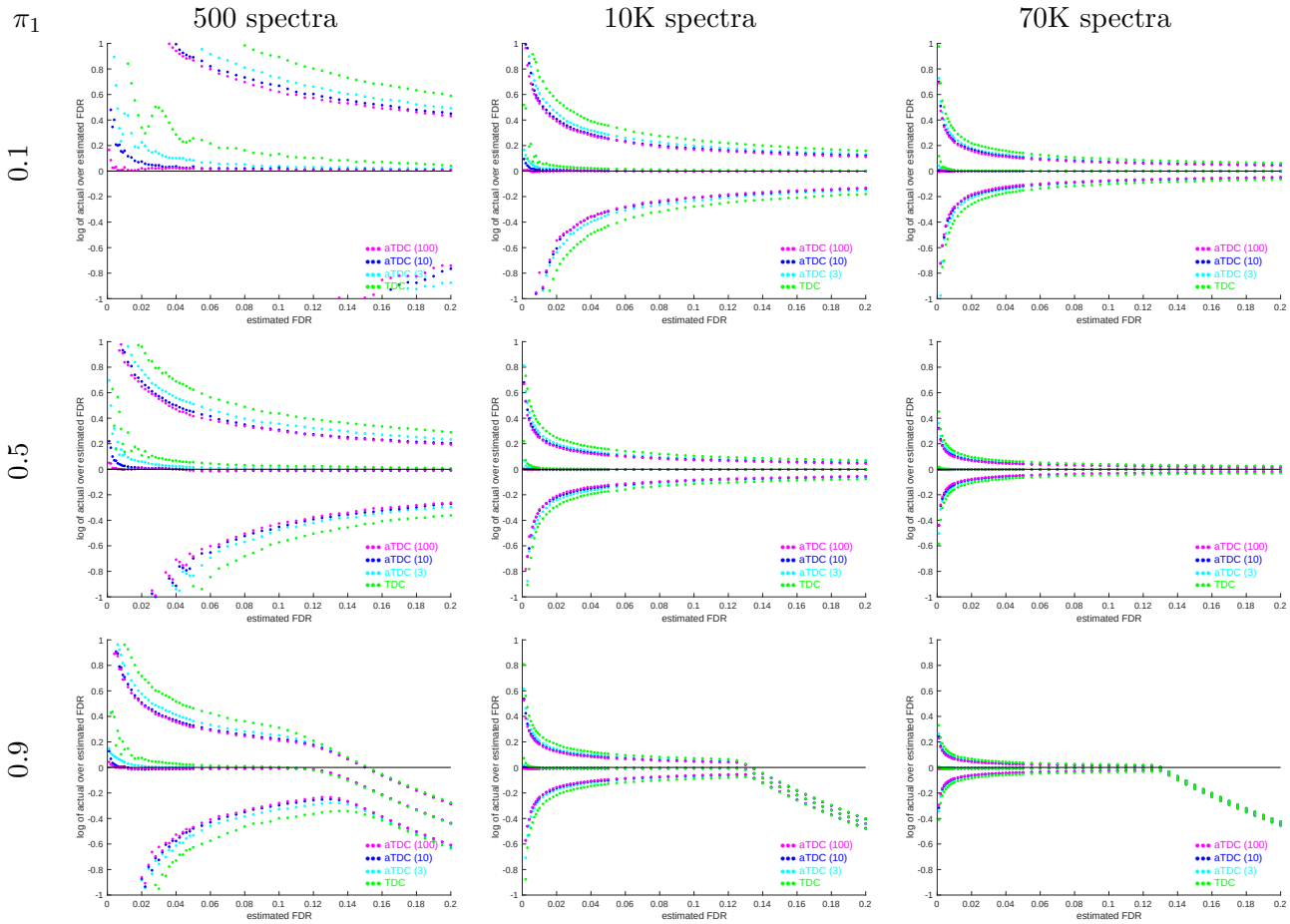


Figure 3: **FDR control: a-TDC and TDC (calibrated scores).** Plotted are the log of the ratios of the mean (empirical FDR, middle curves), 0.05 and 0.95 quantiles (upper and lower curves) of the FDP in the target discovery lists of a-TDC using 1 (TDC), 3, 10, and 100 competing decoys. The liberal bias of TDC is largely addressed by a-TDC, especially using 10 or 100 decoys, and a-TDC exhibits reduced variability in estimating the FDR. Scores are calibrated. 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.

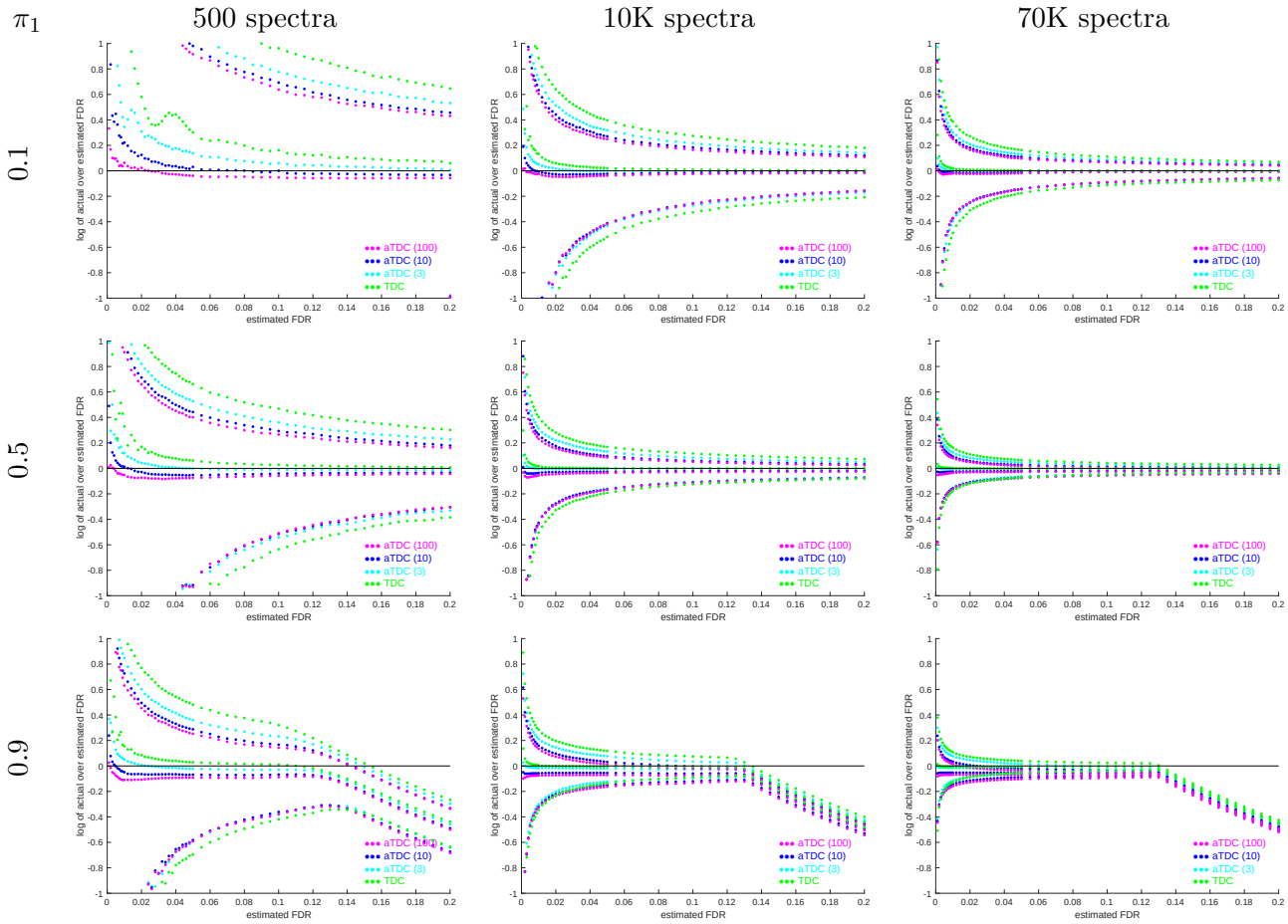


Figure 4: **FDR control: a-TDC and TDC (raw scores).** Similar to Supp. Figure 3 but using uncalibrated scores. While a-TDC still exhibits reduced variability compared with TDC, it also shows a distinct conservative bias that grows with the number of competing decoys. Still, as seen in Supp. Figure 9, in spite of this conservative bias a-TDC typically reports as many as, or slightly more, *correct* discoveries as does TDC.

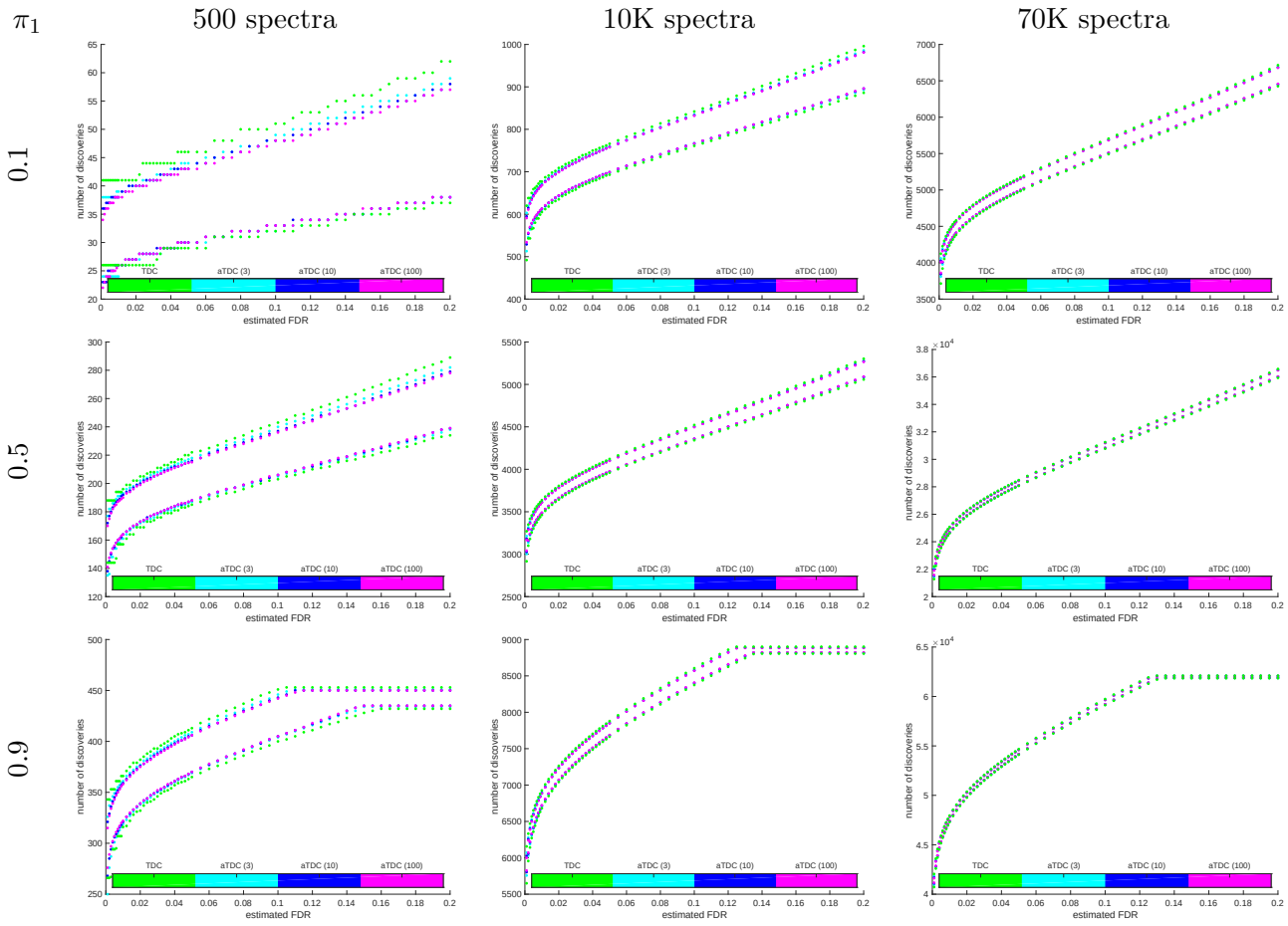


Figure 5: **Variability in number of discoveries: a-TDC vs. TDC (calibrated score).** The 0.05 and 0.95 quantiles of the number of target discoveries demonstrate that, consistent with its reduced variability in estimating the FDR, a-TDC exhibits less variability in the number of target discoveries it 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.

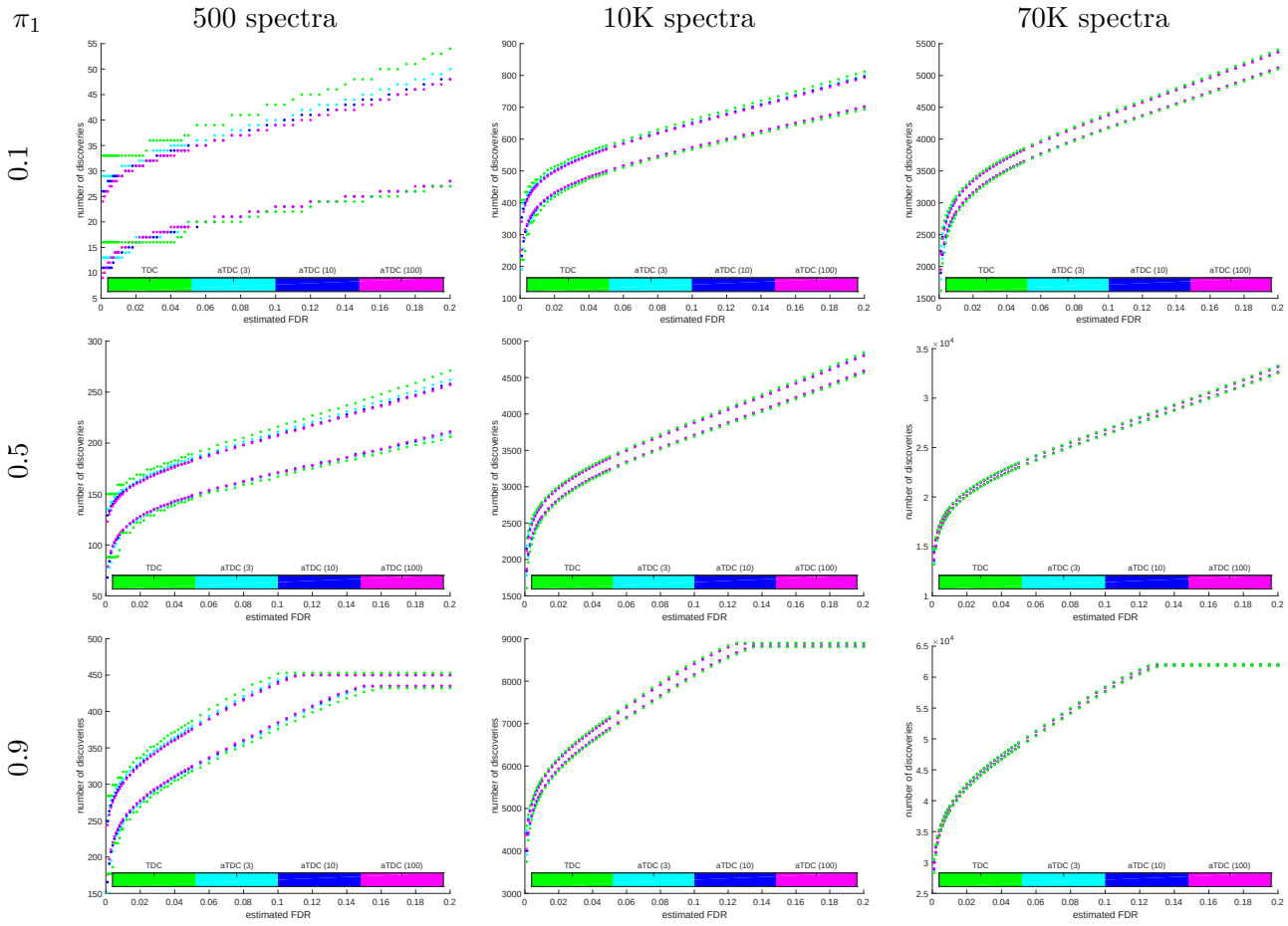


Figure 6: **Variability in number of discoveries: a-TDC vs. TDC (raw scores).** Same as Supp. Figure 5 except using uncalibrated scores. We see that a-TDC is able to even further reduce the variability of TDC. Note that the scale of the $y$-axes vary.

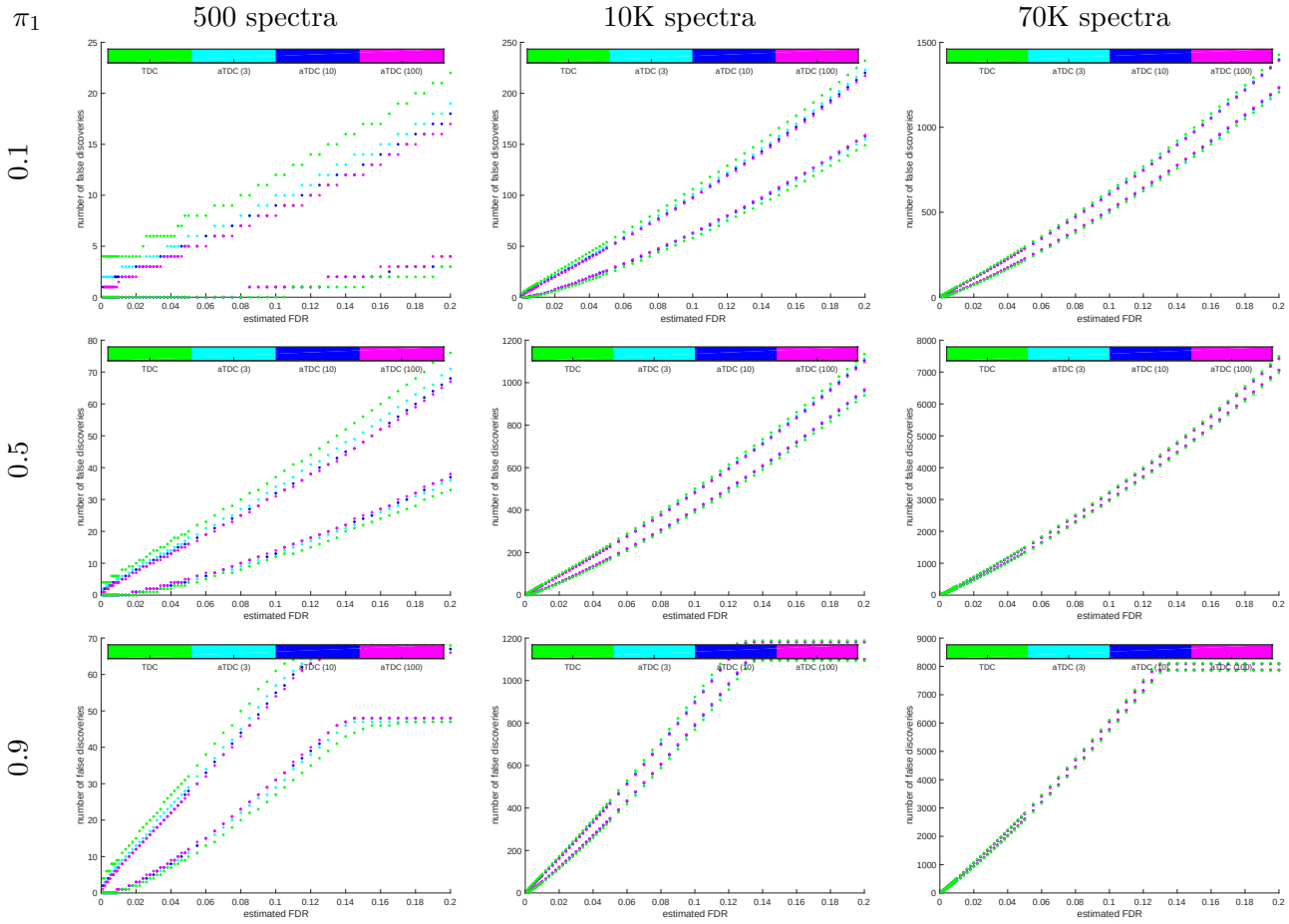


Figure 7: **Variability in number of false discoveries: a-TDC vs. TDC (calibrated score).** The 0.05 and 0.95 quantiles of the number of *false* target discoveries offer a slightly different perspective on a-TDC's reduced variability. All quantiles are taken with respect to 10K simulation runs using 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.

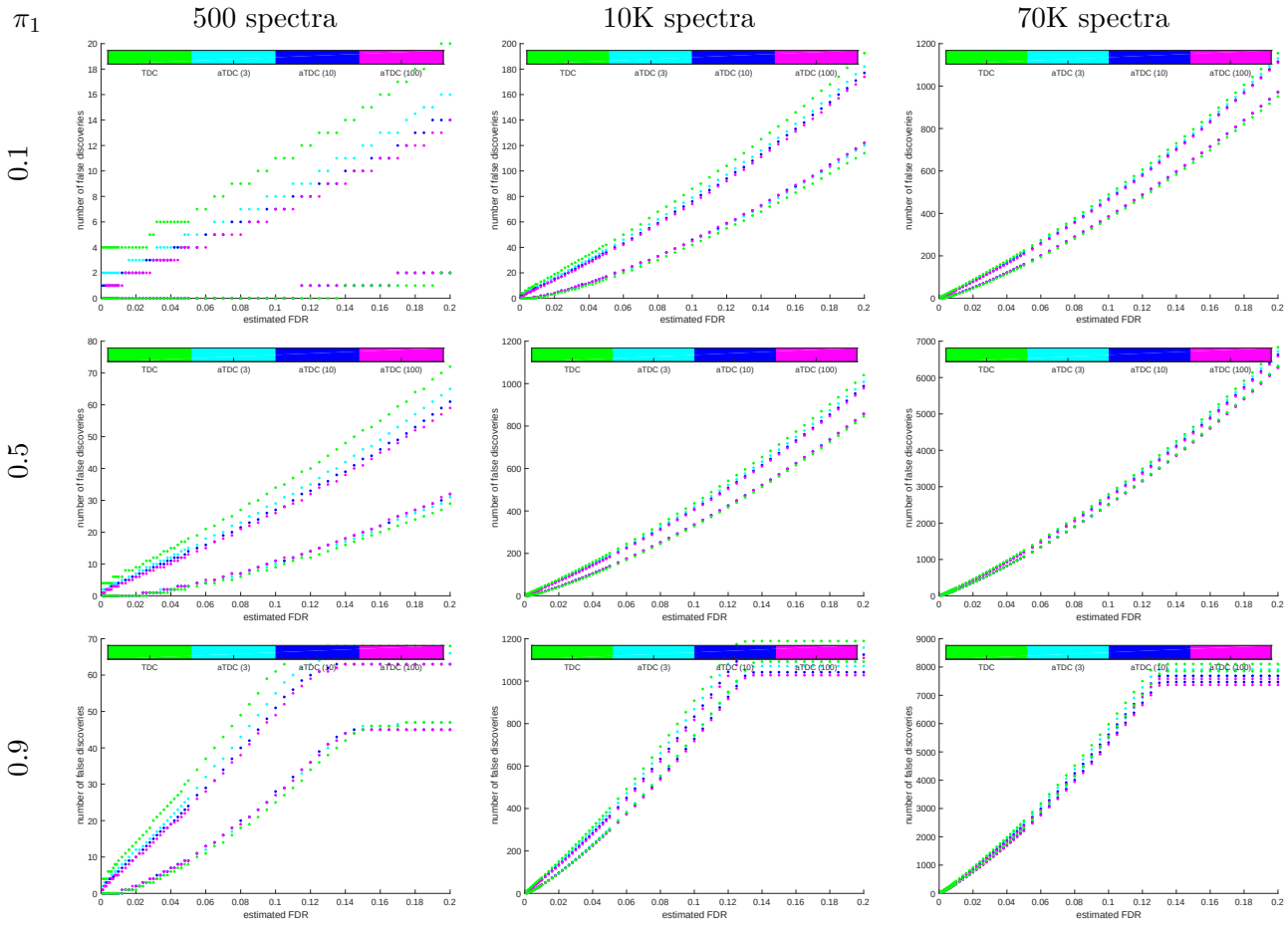


Figure 8: **Variability in number of false discoveries: a-TDC vs. TDC (raw scores).** Same as Supp. Figure 7 only with uncalibrated scores. Again we see that a-TDC is able to reduce the variability of TDC even further than when using calibrated scores.

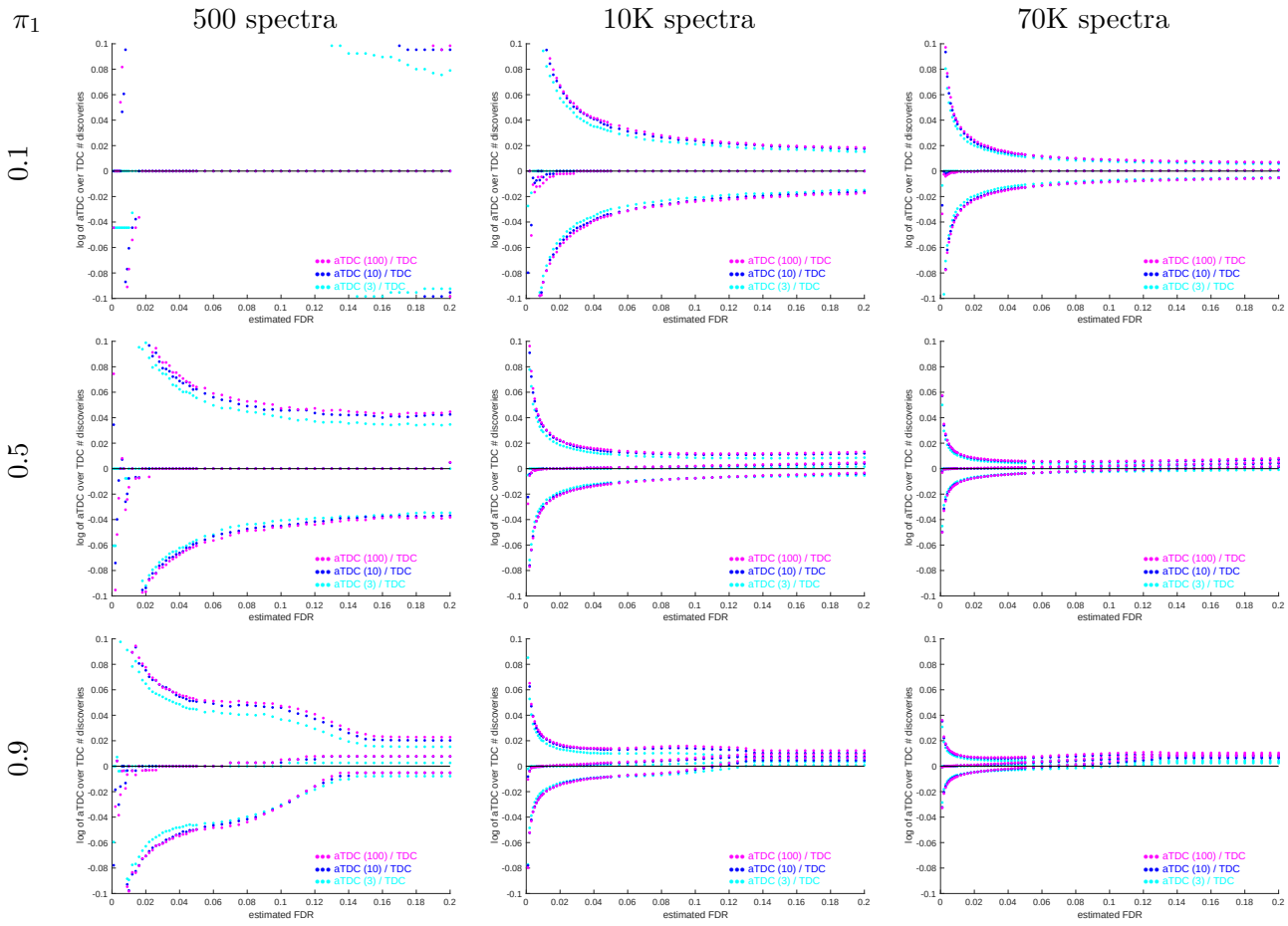


Figure 9: **Power analysis: ratio of number of a-TDC to TDC *true* discoveries (raw scores).** All quantiles are taken with respect to 10K simulation runs using our raw score with 500, 10K, or 70K spectra, 10%, 50%, or 90% of which are native, and 100 candidate peptides per spectrum. The number of competing decoys was 1 (TDC), 3, 10, and 100

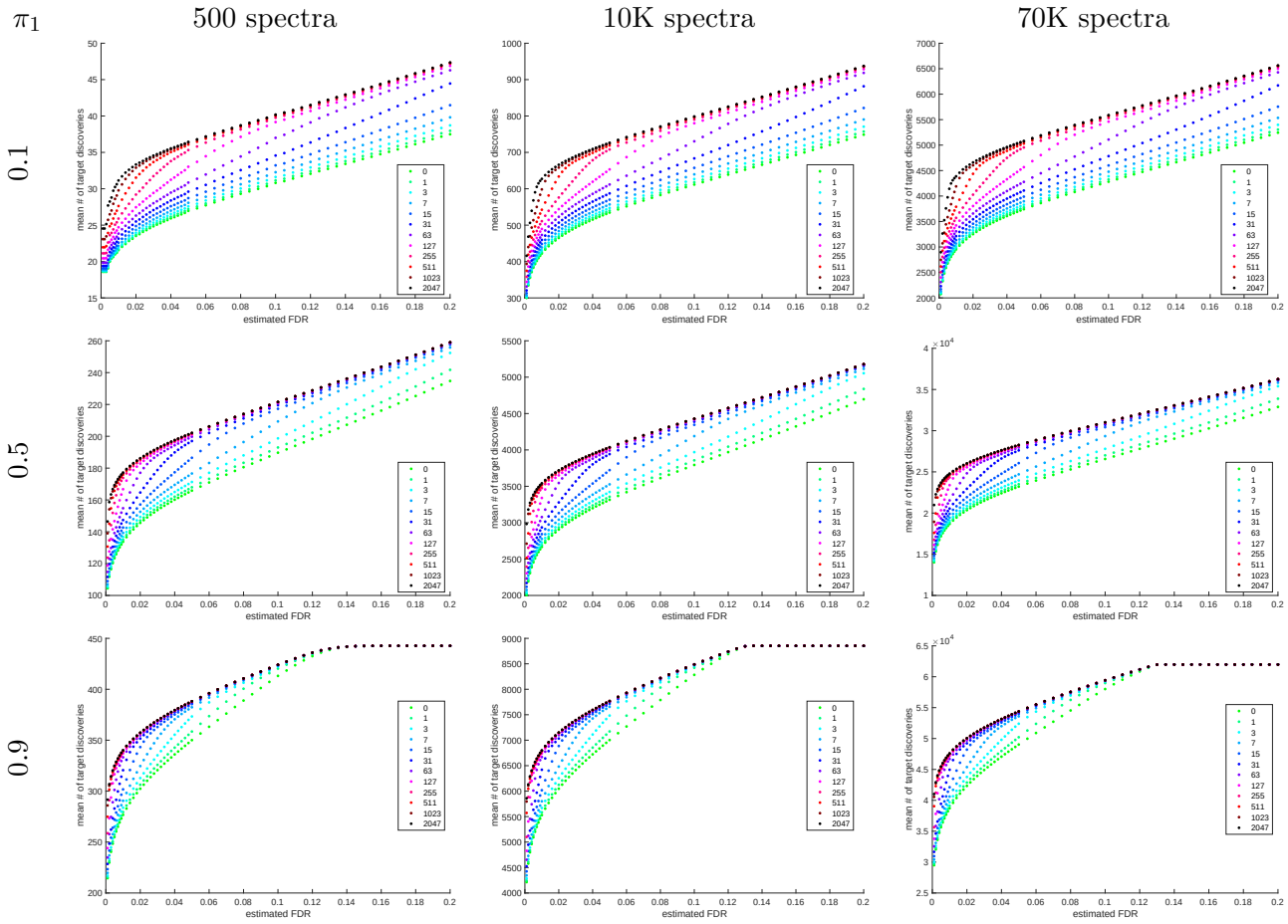


Figure 10: **Partial calibration: Mean number of a-TDC target discoveries.** The panels show the consistent increase in the mean number of a-TDC target discoveries as the number of calibrating decoys increases. All means are taken with respect to 10K simulation runs using our raw score with 500, 10K, or 70K spectra, 10%, 50%, or 90% of which are native, and 100 candidate peptides per spectrum. The number of calibrating decoys was varied from 0 to 2047, and a-TDC uses 10 competing decoys.

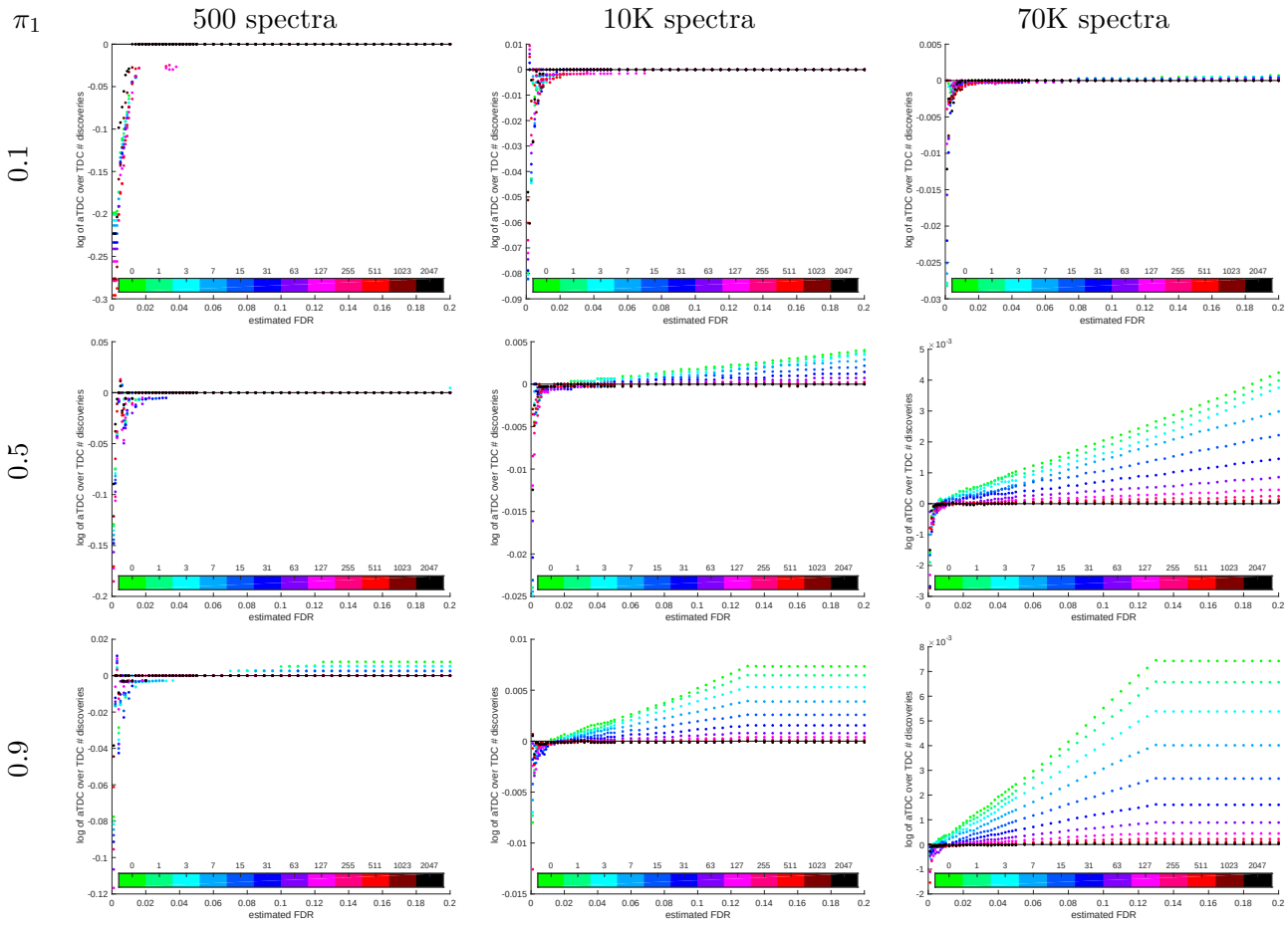


Figure 11: **Partial calibration: aTDC to TDC discoveries.** The power advantage of a-TDC over TDC diminishes with the increase in the number of calibrating decoys from 0 to 2047. Specifically, we see that the log of the median of the ratio of the number of *true* a-TDC discoveries to true TDC discoveries goes to 0 except for very small estimated FDR levels where we know TDC is liberally biased. All medians 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, and a-TDC is using 10 competing decoys.

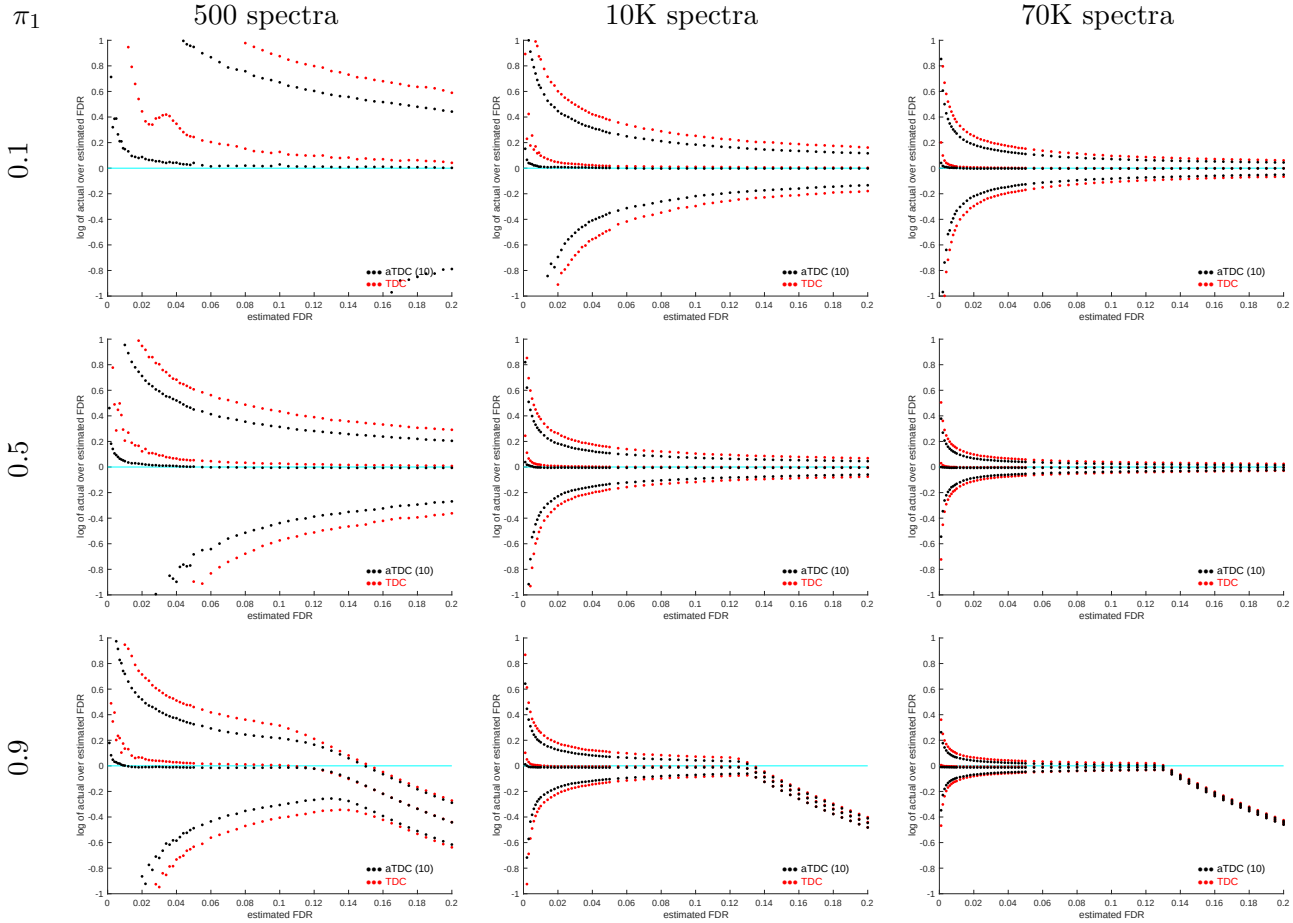


Figure 12: **Partial calibration: a-TDC exhibits reduced variance in FDR estimation.** When combined with partial calibration the a-TDC estimate of the FDR still exhibits smaller variability than TDC's. Plotted are the logs of the mean (empirical FDR), 0.5, and 0.95 quantiles of the actual FDP over the nominal FDR level. The scores here were partially calibrated using 63 calibrating decoys; however, the results for 0 (raw score) as well as for 2047 calibrating decoys were qualitatively the same. 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, and a-TDC is using 10 competing decoys.

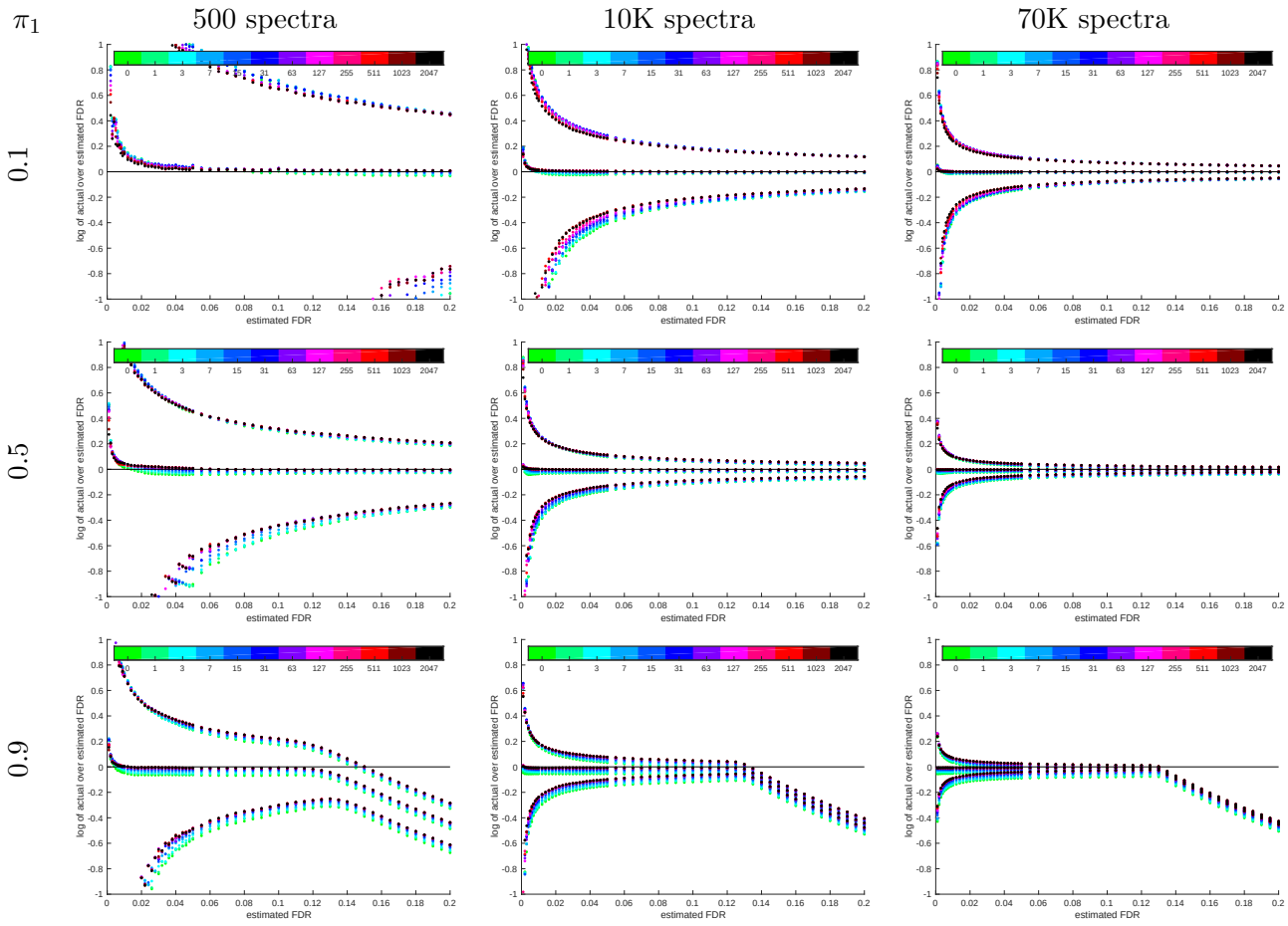


Figure 13: **Partial calibration: FDR control (aTDC).** Same as Supp. Figure 2 except using a-TDC with 10 competing decoys. As the number of calibrating decoys increases from 0 to 2047 a-TDC becomes less conservative: the middle set of curves showing the log of the empirical FDR (mean of FDP across 10K samples) over the nominal level increase toward 0 for small FDR levels. For reference, the log of the 0.05 and 0.95 quantiles (across the same 10K samples) of the FDP over the nominal FDR level are also provided.

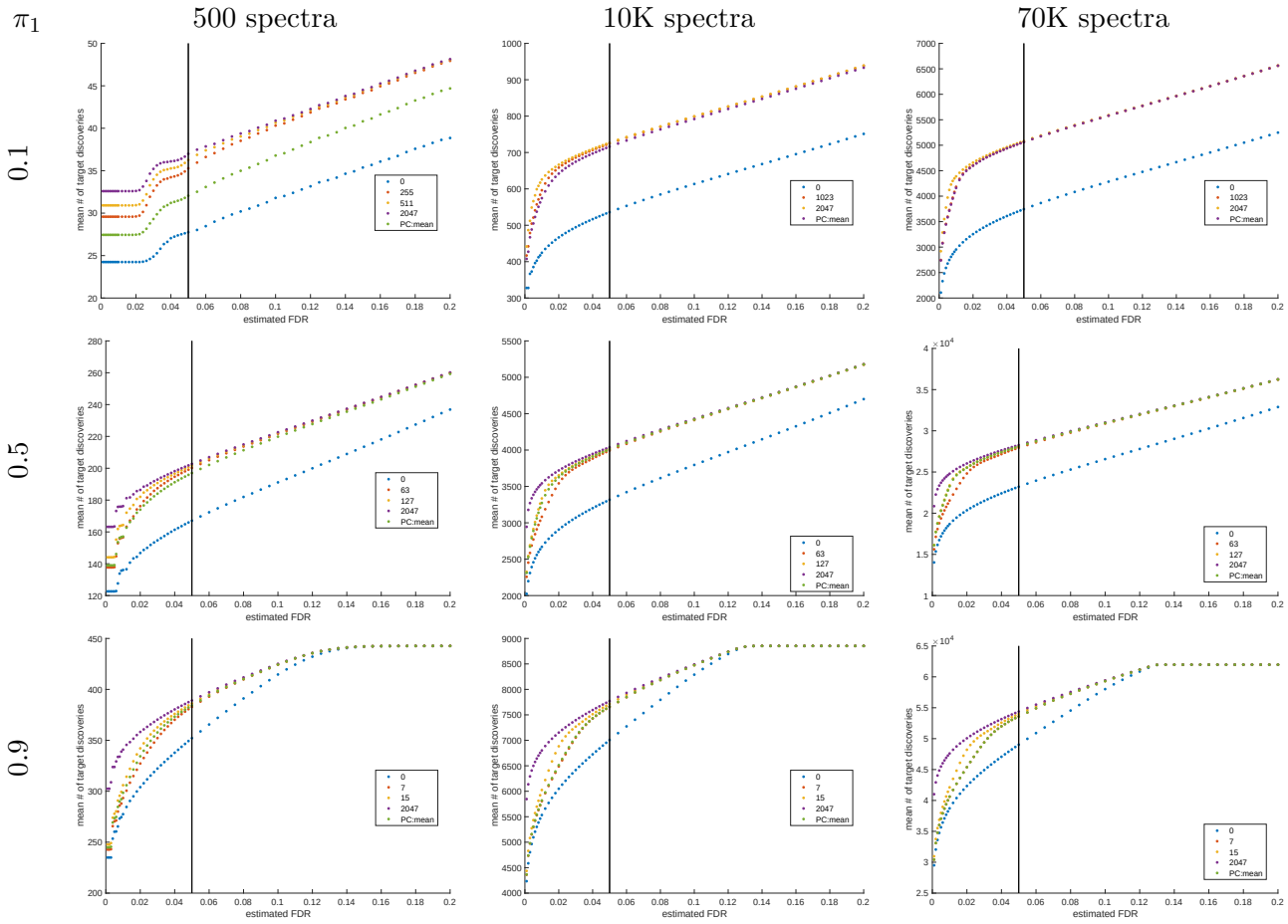


Figure 14: **TDC and progressive calibration (PC).** The panels give the number of TDC discoveries at the nominal FDR levels using: 0 (raw score), $m_1$, $m_2$, 2047 calibrating decoys as well as the number determined by PC. Note that $m_1$ and $m_2$ are the number of decoys in the two cycles that bound the mean of the number of decoys used by PC in the experiment, and they are provided for reference. The vertical bars are located at 0.05, the minimal FDR level of interest for PC in this setup. All means and quantiles are taken with respect to 10K simulations using raw scores, each with 10K spectra, 50% native, and 100 candidate peptides per spectrum.

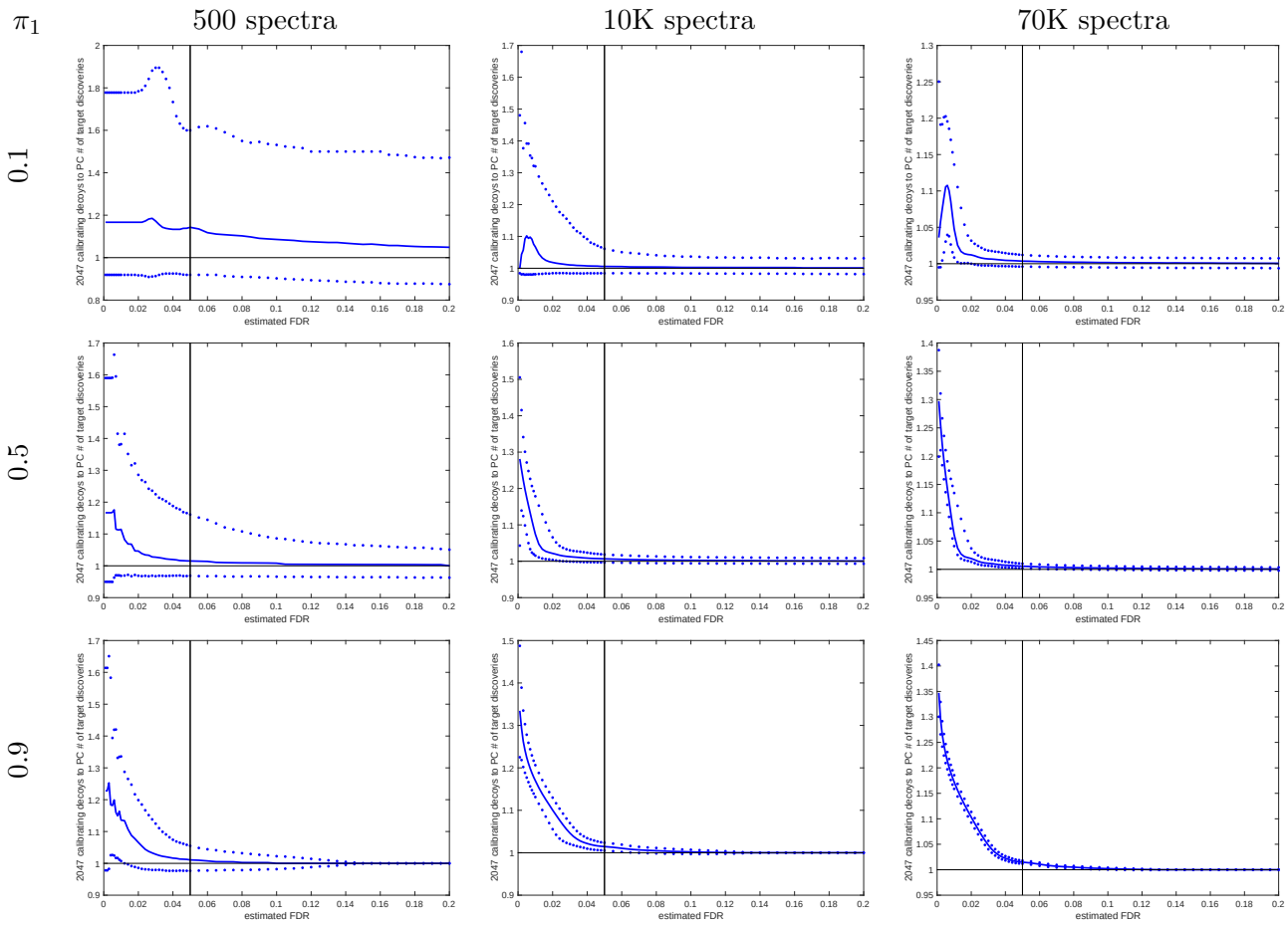


Figure 15: **PC vs. maximal number of calibrating decoys (TDC).** The panels denote the 0.05, 0.5, and 0.95 quantiles of the ratio of the number of TDC discoveries when using the maximal number of 2047 calibrating decoys to the number of discoveries found by PC (tuned to estimated FDR levels $\geq 0.05$). The vertical bars are located at 0.05, the minimal FDR level of interest for PC in this setup. All means and quantiles are taken with respect to 10K simulations using raw scores, each with 10K spectra, 50% native, and 100 candidate peptides per spectrum.

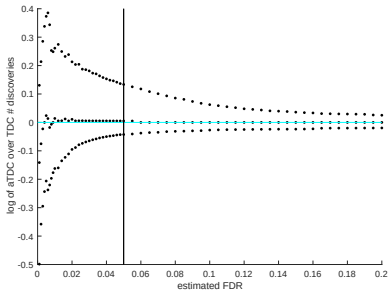


Figure 16: The same quantiles of the ratio of a-TDC (10 competing decoys) to TDC *correct* discoveries (both with PC). The vertical bars are located at 0.05, the minimal FDR level of interest for PC in this setup. All means and quantiles are taken with respect to 10K simulations using raw scores, each with 10K spectra in A-B, and 500 spectra in C, 50% native spectra in all.

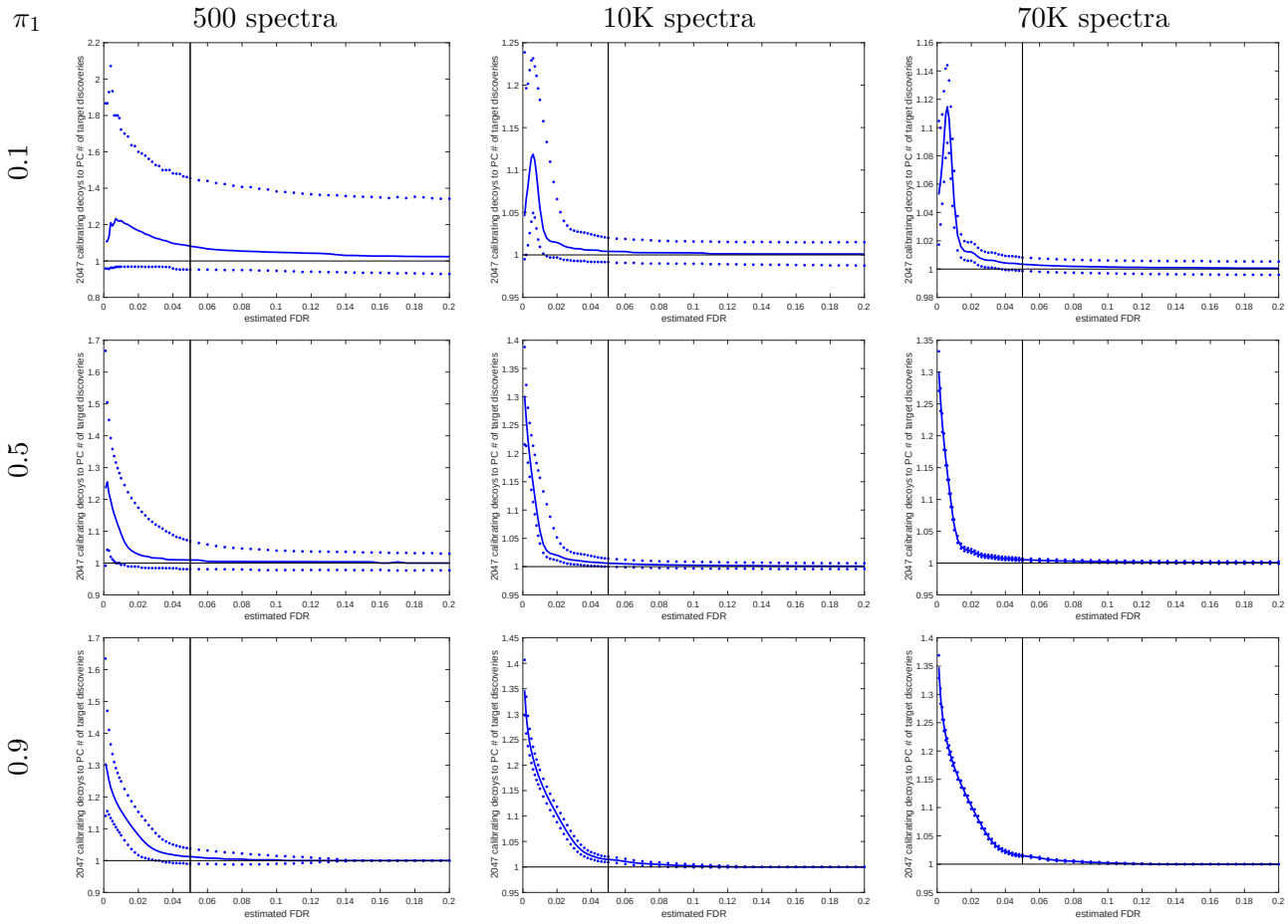


Figure 17: **PC vs. maximal number of calibrating decoys (aTDC).** Same as Figure 15 except PC is coupled with a-TDC using 10 competing decoys.

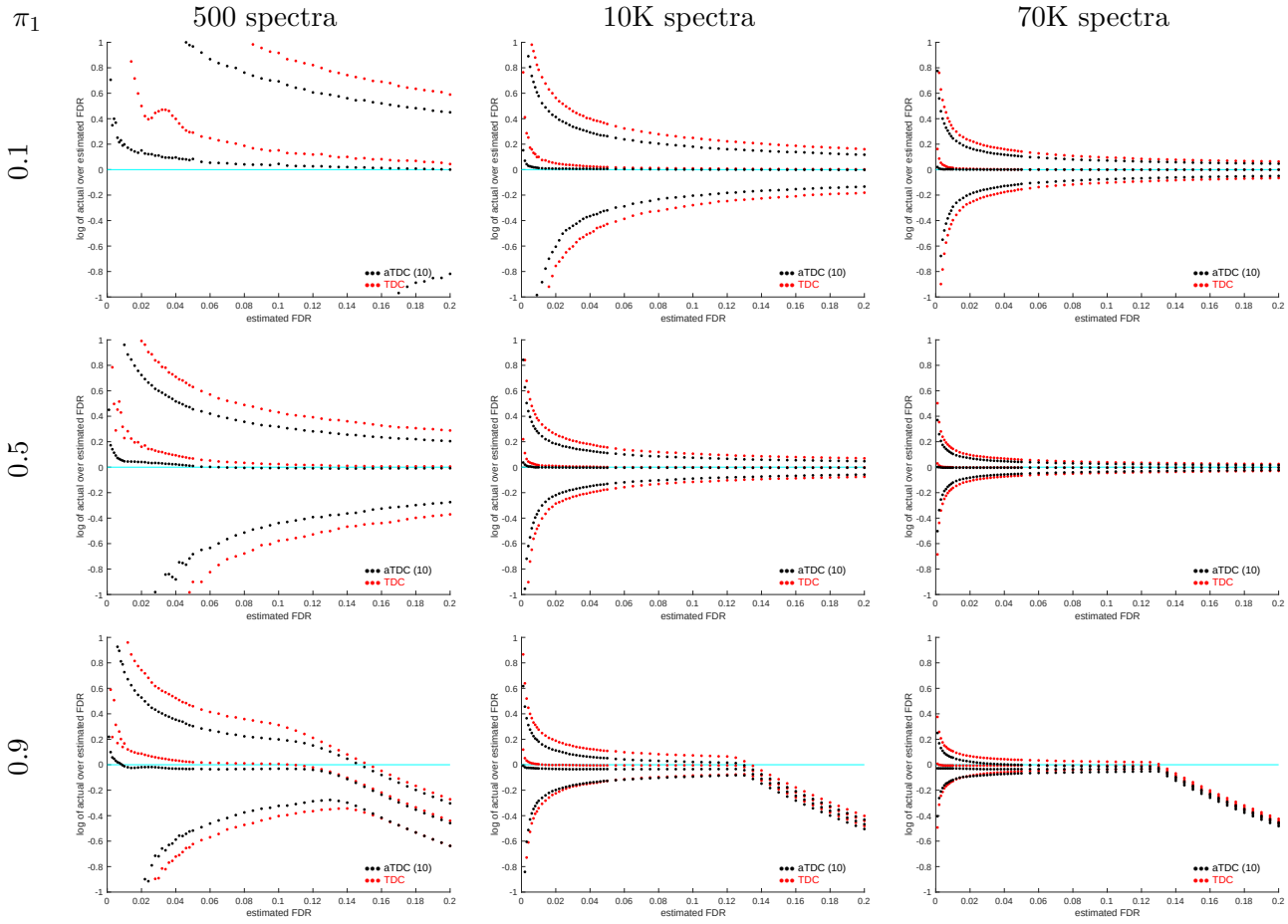


Figure 18: **PC: aTDC's reduced variance.** When combined with PC a-TDC still exhibits smaller variability than TDC in estimating the FDR: plotted are the log of the ratio of the empirical FDR (mean of FDP) as well as the 0.05, and 0.95 quantiles of the FDP to the nominal FDR level. All quantiles and means are taken with respect to 10K samples, and a-TDC is using 10 competing decoys.

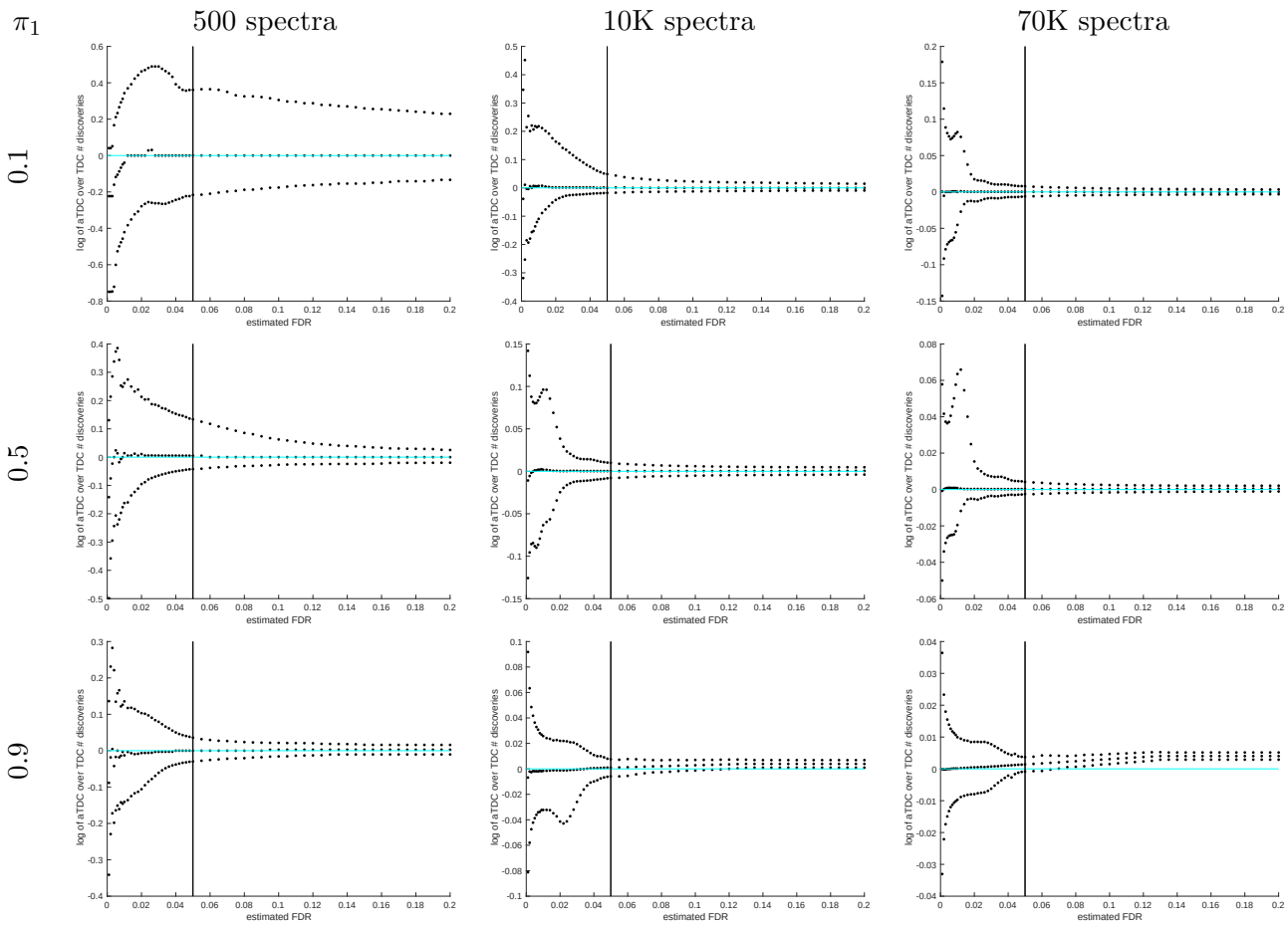


Figure 19: **PC: aTDC's vs. TDC discoveries.** Combined with PC a-TDC can yield a non-negligible number of additional discoveries compared with combining PC with TDC: plotted are the quantiles of the ratio of a-TDC (10 competing decoys) to TDC *correct* discoveries. This increase is due to the fact that a-TDC is less prone to prematurely stopping the doubling procedure. All quantiles are taken with respect to 10K simulations using raw scores.

# 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.

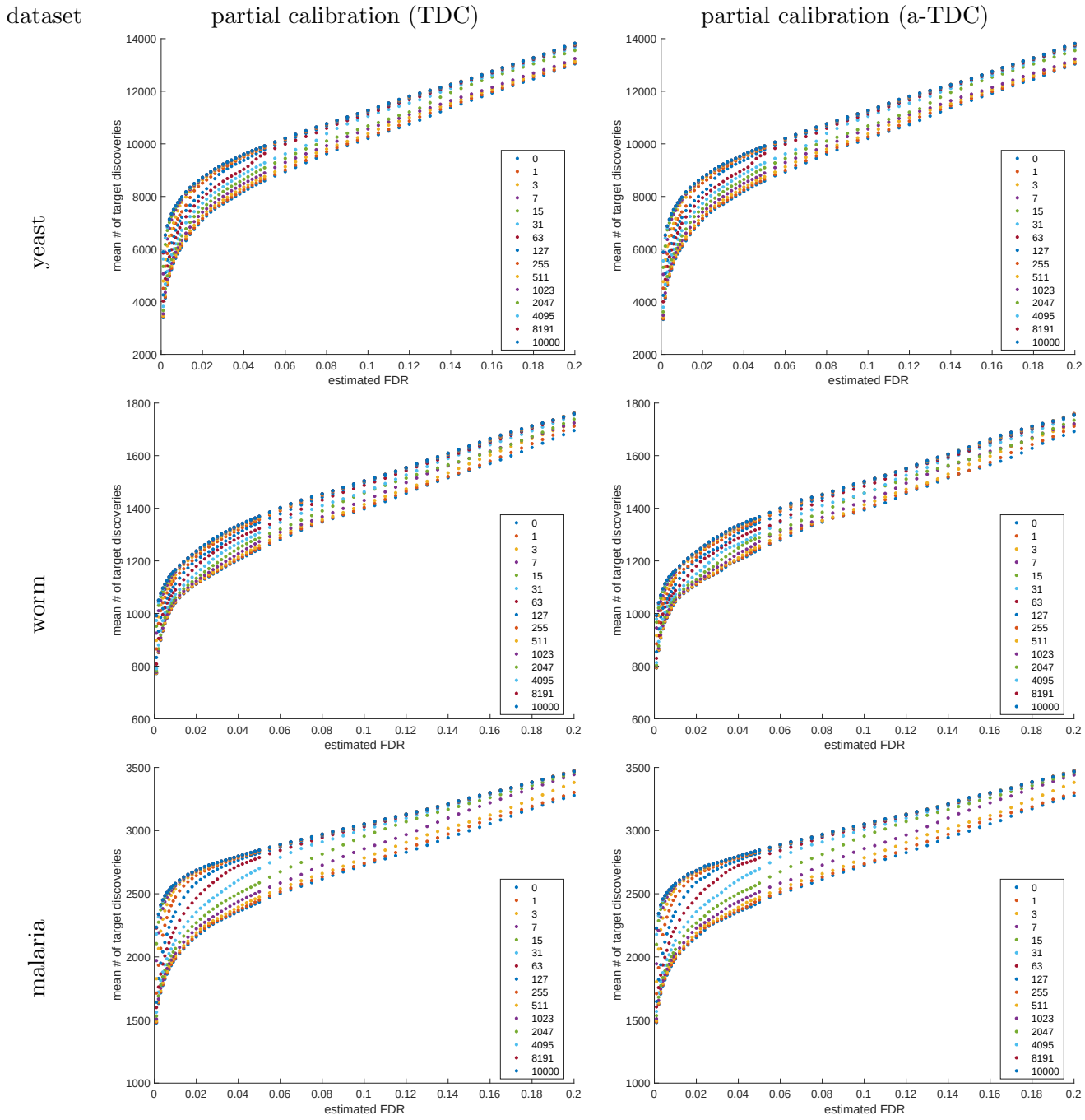


Figure 20: **Real data: partial calibration increases the number of discoveries.** Whether we use TDC (left column) or a-TDC (right column) to control the FDR, the mean number of target discoveries in the real dataset gradually increases with the number of calibrating decoys. All means are taken with respect to 2000 randomly drawn sets of competing decoys (10 for a-TDC, 1 for TDC). The number of calibrating decoys increased from 0 to 10K, and they were selected in an order that was randomly permuted in each of the 2K experiments. See Supp. Section 1.4 for details.

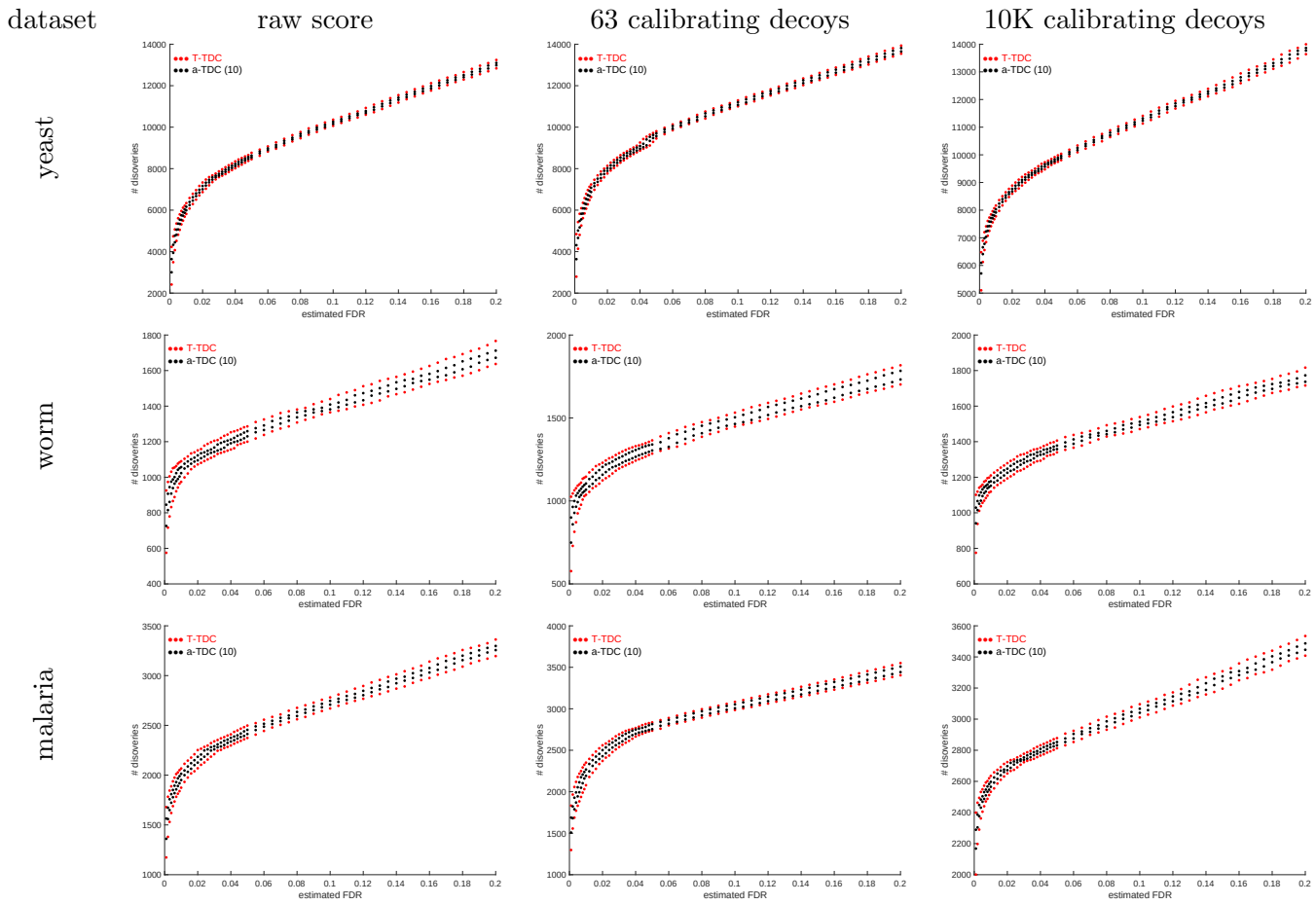


Figure 21: **Real data: a-TDC's reduced variability.** Regardless of the amount of calibration applied to the score, a-TDC is consistently less variable than TDC: the 0.05 and 0.95 quantiles of the number of a-TDC (10 competing decoys) and TDC discoveries are compared in all three sets while increasing the number of calibrating decoys from 0 (raw score), through 63, to the maximal number of 10K. The quantiles are taken with respect to the 2000 randomly drawn sets described in Supp. Figure 20.

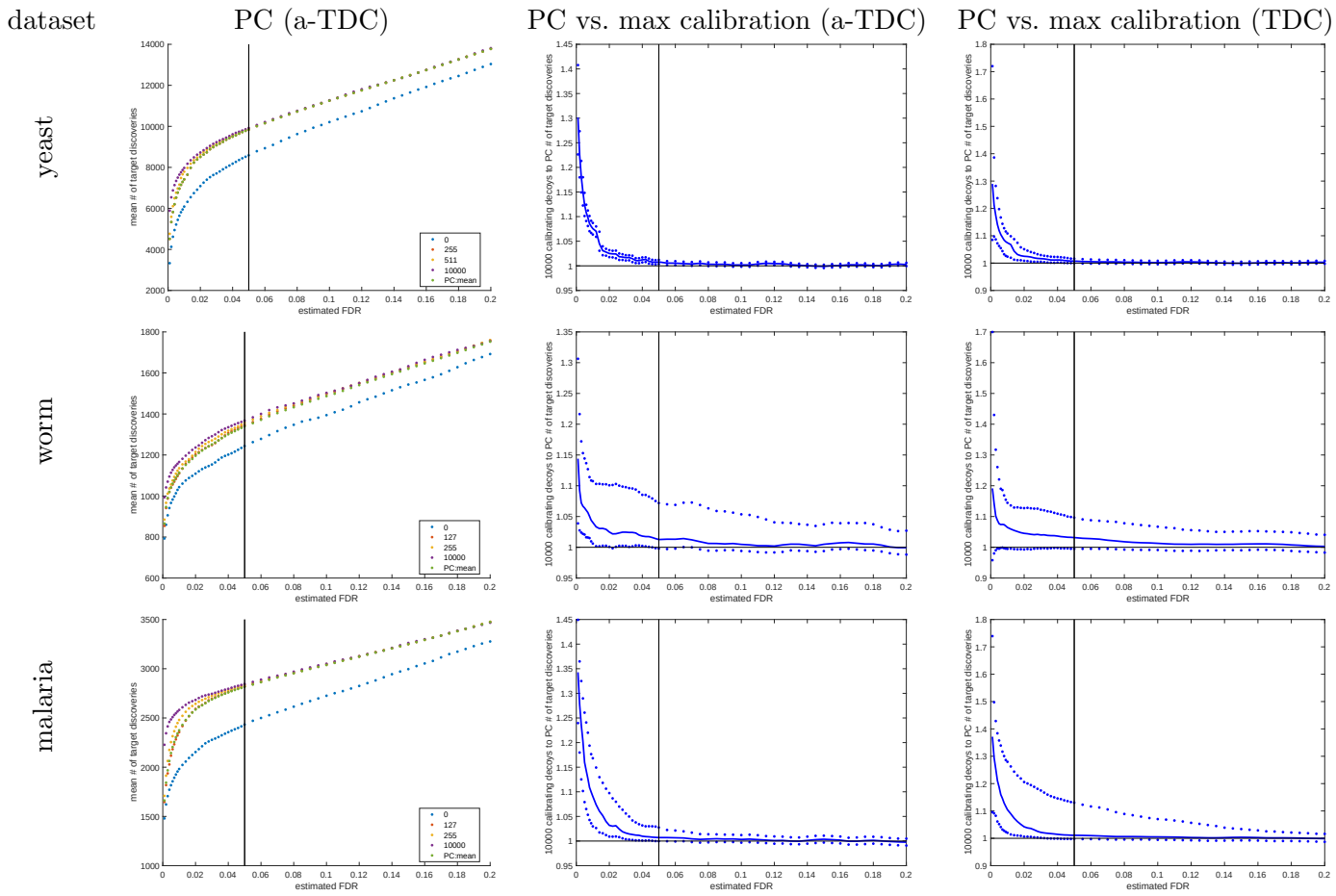


Figure 22: **Real data: progressive calibration.** The left column gives the number of a-TDC discoveries at the nominal FDR levels using: 0 (raw score), $m_1$, $m_2$, 10K calibrating decoys as well as the number determined by PC. Note that $m_1$ and $m_2$ are the number of decoys in the two cycles that bound the mean of the number of decoys used by PC in the experiment (from top to bottom: 262, 235, and 165), and they are provided for reference. The next two columns give a more detailed comparison with the number of a-TDC (middle) and TDC (right) discoveries when using the maximal number of 10K calibrating decoys: plotted are the 0.05, 0.5, and 0.95 quantiles of the ratio of the number discoveries using maximal calibration to the corresponding number using PC. In particular, note that a-TDC typically gives up less discoveries compared with TDC (note that the scale of the $y$-axes varies). The vertical bars are located at 0.05, the minimal FDR level of interest for PC in this setup. The quantiles are taken with respect to the same 2000 randomly drawn sets described in Supp. Figure 20, and in particular a-TDC used 10 competing decoys.