NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS

Faculty of Computer Science
Bachelor's Programme "Data Science and Business Analytics"

Research Project Report on the Topic:

Power Analysis of Various False Discovery Rate Controlling Methods

Submitted by the Student:

group #БПАД222, 3rd year of study — Fokina Viktoriia Konstantinovna

Approved by the Project Supervisor:

Attila Kertesz-Farkas
Professor
Faculty of Computer Science, HSE University

Moscow 2025

Contents

Annotation **3**

1 Introduction **5**

2 Literature overview **6**

3 Methods for FDR Control **7**

3.1 Benjamini-Hochberg procedure 7

3.2 Target-Decoy methods 8

3.2.1 Separated Target-Decoy Methods 9

3.2.2 Combined Target-Decoy Competition (C-TDC) 10

3.2.3 Target-only Target-Decoy Competition (T-TDC) 11

3.3 Mix-Max 11

3.4 Cascaded search 12

4 Results **13**

4.1 Benjamini-Hochberg 13

4.2 Target-decoy competition 14

4.3 STDS, STDS-PIT, T-TDC, C-TDC, MIX-MAX 16

4.4 Cascaded search 18

5 Conclusion **19**

References **21**

Annotation

The study evaluates the statistical power and accuracy of various False Discovery Rate (FDR) controlling methods through simulation analyses. The investigation compares classical approaches (Benjamini-Hochberg procedure) with target-decoy based methods, separated search strategies (STDS, STDS-PIT), competitive approaches (T-TDC, C-TDC), and advanced techniques (Mix-Max). Additionally, cascaded search methodology is investigated as a sequential processing approach. The methods are implemented in simulation with known truth ("ground truth" FDR).

Key findings reveal that STDS is highly conservative while STDS-PIT is slightly anti-conservative, consistent with theory. T-TDC provides asymptotically unbiased FDR estimates, while C-TDC tends to be conservative when applied to target-only discovery lists. Mix-max, which extends STDS-PIT under calibrated scores, achieves nearly unbiased FDR. In contrast, the BH procedure tends to under-estimate FDR in this simulation setting. Notably, the cascaded search strategy achieves the highest number of true discoveries while still maintaining nominal FDR control.

Аннотация

В исследовании оценивается статистическая мощность и точность различных методов контроля коэффициента ложных обнаружений (FDR) с помощью.Сравниваются классические подходы (процедура Беньямини-Хохберга) с методами, основанными на использовании целей-приманок, стратегии раздельного поиска (STDS, STDS-PIT), конкурентные подходы (T-TDC, C-TDC) и усовершенствованные методы (Mix-Max). Кроме того, в качестве подхода к последовательной обработке данных исследуется методология каскадного поиска. Методы реализуются в симуляциях с известным истинным коэффициентом ложных обнаружений.

Основные результаты показывают, что STDS является высококонсервативным, в то время как STDS-PIT несколько антиконсервативен, что соответствует теории. Метод T-TDC обеспечивает асимптотически несмещённые оценки FDR, тогда как C-TDC при применении к наборам данных, содержащих исключительно целевые идентификации, имеет тенденцию к консервативной оценке. Расширение STDS-PIT — Mix-Max — может достигать почти несмещённого контроля FDR. В отличие от этого, процедура BH склонна к недооценке FDR в данной симуляции. Использование каскадной стратегии поиска дает наибольшее число идентификаций при фиксированном номинальном FDR.

Keywords

False Discovery Rate, Statistical Power, Benjamini-Hochberg Procedure, Target-Decoy Competition, STDS, STDS-PIT, Mix-Max, Cascaded Search

1 Introduction

In the era of high-volume data, researchers across various fields such as genomics, proteomics, and mass spectrometry, routinely face the challenge of analyzing thousands or even millions of hypotheses simultaneously. This scenario, known as multiple hypothesis testing, occurs when a statistical analysis involves conducting numerous simultaneous tests for large datasets, each with the potential to yield a meaningful result. When performing such tests, the probability of obtaining false positive results increases substantially, which is known as the multiple testing problem. To guarantee reliable results, error control techniques, which enable preserving the ability to detect true effects, are required.

Historically, the most prevalent approach to error control focused on the Family-Wise Error Rate (FWER), which aims to limit the probability of having at least one false discovery. Such methods as the Bonferroni correction[], for example, adjust thresholds for significance by taking the desired error rate and dividing it by the number of tests performed. In controlling for false positives effectively, such methods become too constraining in high-dimensional data, with considerable loss in statistical power — the ability to discover real signals.

The False Discovery Rate (FDR), introduced by Benjamini and Hochberg in 1995, offers a more balanced approach to multiple testing correction. Instead of controlling for any single false discovery, like FWER, the FDR approach deals with the expected proportion of false discoveries out of all significant findings. For example, a 5% FDR level would mean that, on average, no more than 5% of reported discoveries will represent false positives. This metric has gained widespread adoption because of the ability to maintain statistical power while providing meaningful control over false discoveries.

FDR controlling methods have diversified into various approaches, that can broadly fall under theoretical, empirical, and integrated categories. Theoretical approaches, such as the Benjamini-Hochberg procedure, adjust p-values according to statistical assumptions regarding distribution under the null hypothesis. Despite being computationally efficient and widely adopted in software packages, it relies heavily on such assumptions as uniformity of p-values under the null hypothesis or independence — conditions not always encountered in real datasets. Empirical methods, such as Target-Decoy Competition (TDC) introduced by Elias and Gygi, address these limitations through use of synthetic "decoy" datasets — artificially generated null instances — to directly estimate false discoveries in the data. For example, in proteomic studies, decoy sequences can mimic reversed peptide sequences, which allows researchers to approximate the number of false matches in the results. Decoy-based methods avoid theoretical assumptions, however their

accuracy is based on the quality and quantity of decoy samples. Integrated approaches, such as STDS-PIT and Mix-Max, aim to combine theoretical adjustments with empirical data. These methods use decoy-based empirical distributions and apply theoretical corrections to adjust for factors like the proportion of true null hypotheses.

The aim of this study is to systematically evaluate and compare the performance of these FDR controlling approaches through the simulation analyses. By testing these methods under controlled conditions the goal is to identify the trade-offs between statistical power and FDR control accuracy.

2 Literature overview

The false discovery rate (FDR) concept, introduced by Benjamini and Hochberg (1995) [1] as the expected proportion of false positives among all rejected hypotheses, provides a less conservative error criterion than family-wise error rate (FWER). The BH procedure orders p-values and rejects hypotheses up to a chosen FDR level, which greatly increases statistical power in large-scale testing scenarios. John D. Storey (2002) [6] extended this framework by introducing the q-value, a data-driven analogue of the p-value for FDR control, and proposed to estimate the proportion of true null hypotheses. In Storey's approach, one fixes a rejection region and then estimates the resulting FDR, rather than fixing FDR and finding a threshold. This q-value methodology often yields much higher power than the original BH step-up method.

In fields like proteomics, empirical FDR estimation has become popular. Elias and Gygi (2007) [2] proposed the target-decoy competition (TDC) method for peptide identification: each spectrum is searched against a concatenated database of "target" (real) and "decoy" (shuffled or reversed) sequences, and only the higher-scoring match of each target–decoy pair is retained. The FDR is then estimated from the fraction of retained decoy matches. This approach is intuitive and widely used, but has known limitations. In particular, random decoys can occasionally outscore true targets, causing some valid identifications to be dropped and the estimated FDR can vary with the decoy generation. Variants such as the "target-only" TDC attempt to mitigate these issues, but all empirical decoy-based methods rely on well-calibrated scoring and can either over- or under-estimate the true error rate under realistic conditions.

Käll et al. (2008) [3] introduced the STDS-PIT method: after a separate target–decoy search (STDS), one estimates the overall fraction of false targets (the "percentage of incorrect targets" or PIT) and multiplies the naive decoy-based FDR by this PIT factor. This adjustment boosts sensitivity by accounting for the fact that not all target-spectrum matches are correct.

STDS-PIT assumes well-calibrated scores and can be anti-conservative (underestimate the true FDR) in practice. More recently, Keich et al. (2015) [4] proposed the Mix-Max procedure as an extension of STDS-PIT. Mix-max assumes calibrated score distributions and models each spectrum as "native" or "foreign" using mixture modeling to estimate the expected number of false discoveries from both groups. It retains all high-scoring target matches while avoiding biases of pure target-decoy competition.

3 Methods for FDR Control

Assume m null hypotheses are being tested. Out of these, m_0 are truly null and $m_1 = m$ - m_0 are truly alternative. The outcomes can be categorized as follows:

	Declared non-significant	Declared significant	Total
True null	TN	FP	m_0
True alternative	FN	TP	m_1
Total	m-R	R	m

Where:

- TN: True Negatives
- FP: False Positives
- FN: False Negatives
- TP: True Positives
- R: Total number of rejected null hypotheses

The False Discovery Rate is defined as

$$\text{FDR} = \mathbb{E}\left[\frac{FP}{R}\,|R > 0\right] \cdot \mathbb{P}(R > 0)$$

3.1 Benjamini-Hochberg procedure

Two distributions were simulated: a null distribution $N(0, 1)$ representing true negatives and a target distribution $N(3, 1)$ representing true positives. Each distribution contained 10,000 observations. This results in a dataset with a true proportion of null hypotheses, π_0, of 0.5.

The p-values were calculated using the right-tailed probability: $\text{P}(\text{X} \geq \text{x}) = 1$ - $\Phi(x)$, where $\Phi(x)$ is the cumulative distribution function of the standard normal distribution. This calculation represents the probability of observing a test statistic as extreme or more extreme than the one observed, under the null hypothesis. On figure [3.2] it is highlighted from which distribution the

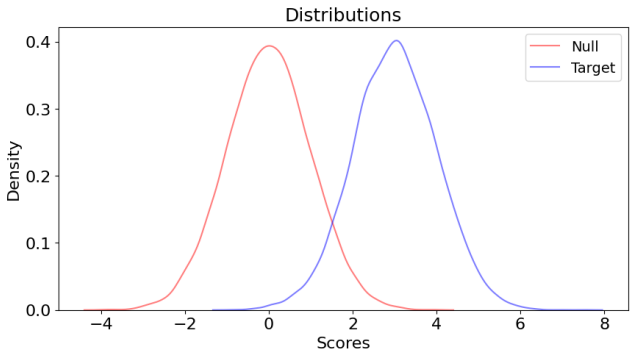

Figure 3.1: Distribution of generated scores

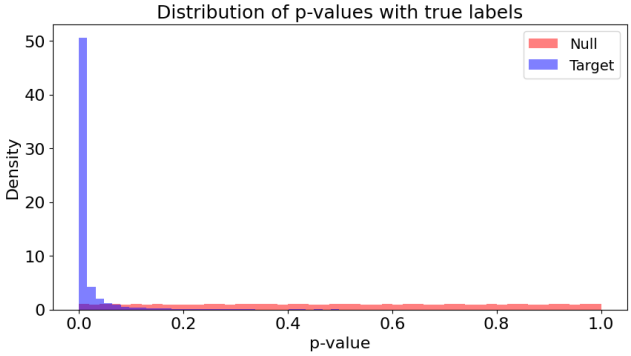

Figure 3.2: p-values distribution

p-values originate. It can be seen that under the null hypothesis p-values are uniform and for the target distribution most values are shifted to the right (mean 3) and therefore most of them will get very small p-values.

The classical Benjamini-Hochberg procedure for controlling the FDR operates on the ordered p-values from m hypothesis tests and works as follows:

1. **Set FDR level**: Choose a significance level α (e.g., $\alpha = 0.05$).

2. **Order p-values**: Arrange the p-values $p_1, p_2, \ldots, p_m$ from m hypothesis tests in ascending order: $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$

3. **Calculate thresholds**: For each rank i (where $i = 1, 2, \ldots, m$), calculate threshold: $\frac{i \cdot \alpha}{m}$

4. **Find largest significant p-value**: Find the largest i such that: $p_{(i)} \leq \frac{i \cdot \alpha}{m}$

5. **Reject null hypotheses**: Reject the null hypotheses for all tests with p-values $p_{(1)}, p_{(2)}, \ldots, p_{(i)}$.

The Benjamini-Hochberg procedure establishes a critical threshold line y = $\frac{i \cdot \alpha}{m}$, where i is the rank of a p-value when sorted in ascending order, m is the total number of tests, and α is the desired false discovery rate (FDR). This dynamic thresholding prefers first to detect strong signals (small p-values) early and then increasingly widen acceptance in a manner that permits weaker signals, but not over the bounded FDR.

3.2 Target-Decoy methods

The Benjamini–Hochberg procedure guarantees FDR control under the key assumption that p-values from true nulls are uniformly distributed on [0, 1]. In practice—especially in fields like computational biology this assumption can fail because of complex dependencies, unusual test statistics, or unknown null distributions.

The Target-Decoy approach is an alternative and widely used method for controlling the False Discovery Rate in fields like proteomics and genomics. The technique involves creating a

"decoy" set of data that mimics the structure of the target data but is known to contain no true signals. By comparing results from the target and decoy datasets, the FDR can be estimated empirically without relying on theoretical assumptions about the distribution of p-values.

For the following methods the simulation captures key features of mass spectrometry by distinguishing two types of spectra:

Native Spectra: These correspond to spectra generated by peptides that are actually present in the target database. Each native spectrum has a true match with a score $X_i \sim \mathcal{N}(2.5, 1)$, but this match competes with random incorrect matches drawn from the remainder of the database, where $Y_i \sim \mathcal{N}(0, 1)$. The observed target score is defined as $W_i = \max(X_i, Y_i)$. As a result, even native spectra can be misidentified if a random match outperforms the true one.

Foreign Spectra: These represent spectra from peptides absent in the target database, such as contaminants, modified peptides, or unrelated ones. In this case, no valid match exists, so $X_i = -\infty$, and the observed target score becomes $W_i = Y_i \sim \mathcal{N}(0, 1)$, reflecting a random match that is always incorrect.

Decoy Scores: Each spectrum, whether native or foreign, is also assigned a decoy score $Z_i \sim \mathcal{N}(0, 1)$, drawn independently from a decoy database. These scores provide a baseline estimate of the null distribution and are used in target-decoy competition for FDR estimation.

π_0 is set to 0.5, meaning that half of the spectra are foreign and half are native.

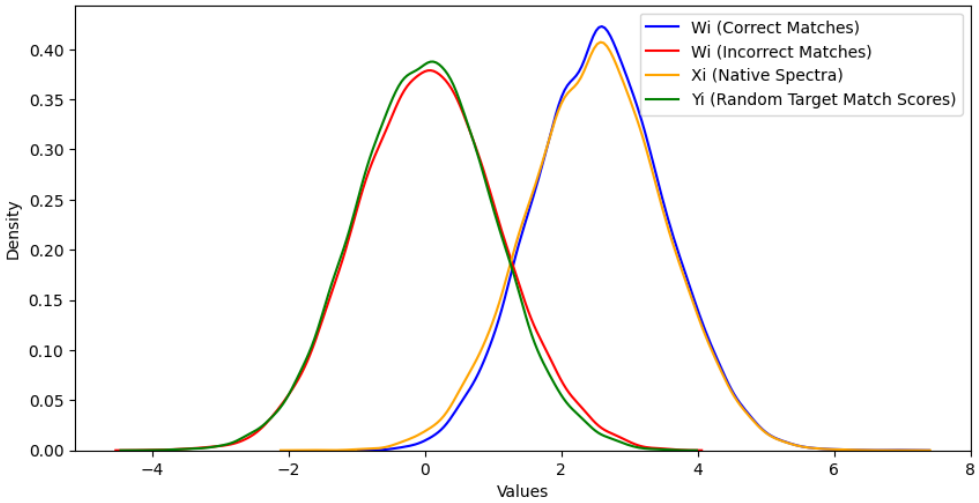

Figure 3.3: Target-Decoy distributions

3.2.1 Separated Target-Decoy Methods

STDS is an approach to empirical FDR estimation that treats target and decoy searches as independent processes. Introduced by Käll et al., it was designed to avoid the issue in competitive

methods where true high-scoring target observations can be lost due to random decoy wins.

In STDS, each observation is tested separately against target and decoy distributions. In this simulation decoy samples are generated from the null distribution. Since these decoy matches occur by random chance, they provide an empirical null distribution that estimates how many random high scores would occur among truly null observations.

STDS assumes that incorrect matches to the target database happen at the same rate as matches to the decoy database. This means the number of decoy matches can be used to estimate the number of false positives in the target matches.

Every observation yields two scores: one from the target database (the best match) and one from a decoy database. To estimate how many identifications could be false, STDS counts how many decoy scores exceed a chosen threshold and divides that number by how many target scores exceed the same threshold, formally defined as

$$\frac{\widehat{F}}{D} = \frac{|\{i : z_i > T\}|}{|\{i : w_i > T\}|}, \tag{1}$$

where T in eq.[1] is the score threshold, z_i are decoy scores, and w_i are observed target scores. There is no head-to-head competition between target and decoy scores.

STDS with Percentage of Incorrect Targets (STDS-PIT) enhances the basic STDS by using an estimate of the fraction of observations that are truly null. While STDS assumes all target discoveries could be incorrect, in reality, many observations genuinely have alternative signals. STDS-PIT recognizes this by incorporating an estimate of π_0 - the proportion of observations that are truly null, defined as

$$\frac{\widehat{F}}{D} = \widehat{\pi}_0 \frac{|\{i : z_i > T\}|}{|\{i : w_i > T\}|}, \tag{2}$$

where π_0 is the estimated proportion of foreign spectra.

Decoy scores are first used to construct an empirical null distribution, which allows for the direct estimation of p-values for target hypotheses by comparing observed scores against the decoy-derived empirical cumulative distribution function.

3.2.2 Combined Target-Decoy Competition (C-TDC)

C-TDC is the original competitive target-decoy method introduced by Elias and Gygi. It estimates FDR in the combined list of target and decoy hits after target-decoy competition and laid the groundwork for later developments in proteomics FDR estimation.

The method assumes that incorrect target and decoy hits are equally likely to win the

competition. Based on this, it treats the number of decoy hits as half the total number of false positives. This leads to the FDR being calculated as twice the number of decoy wins divided by the total number of discoveries (targets and decoys combined).

After each observation's target and decoy scores compete, whichever score is higher is retained. The discovery list includes both target and decoy wins.

Formally, let S = s_1, s_2, ..., s_n represent target scores and D = d_1, d_2, ..., d_n represent decoy scores drawn from the null distribution $N(0, 1)$. For any threshold T, the target-decoy FDR estimator is:

$$\widehat{\text{FDR}}_2(T) = \frac{2\,|\{i : d_i \geq T\}|}{|\{i : s_i \geq T\}| + |\{i : d_i \geq T\}|}. \tag{3}$$

The factor of 2 accounts for decoys representing only half the expected false positives when decoys and targets compete.

3.2.3 Target-only Target-Decoy Competition (T-TDC)

T-TDC is a competitive target-decoy method, but unlike combined methods that report both target and decoy wins, T-TDC keeps only the target wins—the cases where the target score beats the decoy. This filtering mirrors the real-world goal of identifying target peptide identifications, not decoy hits.

The core idea behind T-TDC is that, for incorrect matches, the target and decoy scores are statistically symmetric—each has a 50% chance of winning. This assumption, expressed as $\text{P}(Z_i > Y_i \mid \text{False Match}) = 0.5$, enables unbiased FDR estimation for the filtered list of target discoveries.

For each observation, both a target and decoy scores are computed. Whichever score is higher "wins" the competition and is added to the discovery list. If the decoy wins, it is discarded—but counted in the FDR estimation. The FDR is then estimated at each threshold by taking the ratio of decoy wins to target wins.

$$\widehat{\text{FDR}}_1(T) = \frac{|\{i : d_i \geq T\}|}{|\{i : s_i \geq T\}|}. \tag{4}$$

3.3 Mix-Max

The Mix-Max approach extends both theoretical and empirical frameworks further through partitioning false discovery estimates into two components: a π_0-adjusted null observation contribution (F_0) and and an alternative observation false discovery component (F_1). The first com-

ponent, F_0, uses the estimated proportion of null observations, π_0, to scale the observed decoy counts above a given score threshold, representing false discoveries from observations that are truly null. The second component, F_1, estimates false discoveries among alternative observations where random noise scores higher than true signals, calculated through cumulative distribution ratios that account for the $\max(X_i, Y_i)$ structure of alternative observation scores. By this two-part estimation, Mix-Max effectively distinguishes between false discoveries arising from truly null observations and those resulting from incorrect evaluations among observations that do have true alternative signals.

The false discoveries from foreign spectra are estimated as: $\hat{F_0} = \hat{\pi_0} \times |\{i{:}z_i > T\}|$

This component follows the same logic as STDS-PIT but applies only to the null observation population. Since null observations by definition have no true signals, any discovery above threshold represents a false positive, and the decoy distribution provides an unbiased estimate of this rate.

The false discoveries from alternative observations are calculated:

$$\hat{F_1} = (1 - \hat{\pi_0}) \cdot \sum_{z_j > T} \left[\frac{\sum_k 1_{w_k \leq z_j} - \hat{\pi}_0 \sum_k 1_{z_k \leq z_j}}{(1 - \hat{\pi}_0) n_\Sigma} \Big/ \frac{\sum_k 1_{z_k \leq z_j}}{n_\Sigma} \right]_{[0,1]}$$

The final Mix-Max estimate combines both components: $\hat{F} = \hat{F_0} + \hat{F_1}$

3.4 Cascaded search

Cascaded search [5] is a sequential method that processes datasets in stages, removing identified targets at each step. It processes datasets one at a time, beginning with the one expected to yield the most discoveries. After identifying targets in a dataset, those targets are excluded from all subsequent searches, preventing any double-counting. Availability is tracked dynamically: once an alternative hypothesis is confirmed in any dataset, it is removed from the remaining search pool. This setup mimics real-world workflows, where early successes reduce the number of remaining candidates.

A realistic cascaded search scenario was created with 10 datasets containing decreasing numbers of alternative hypotheses (N_1 samples from $N(3, 1)$) ranging from 3000 in the first dataset to 300 in the final dataset. Each dataset also contains null hypotheses (N_0 samples from $N(0, 1)$) to reach a total size of 20,000 samples per dataset. This structure mimics real-world scenarios where initial searches target high-confidence databases with many true matches, while subsequent searches explore increasingly sparse or lower-quality resources.

For each dataset, Benjamini–Hochberg procedure was applied independently using the true null proportion (π_0) specific to that set. By controlling the false discovery rate (FDR) independently at each step, statistical accuracy is maintained even as the target pool shrinks.

As a benchmark, this cascade method is compared to a "maximum dataset" approach. Here, a single composite dataset is formed by taking the maximum value in each row across the ten original datasets. Empirical p-values were calculated against a null distribution formed by row-wise maxima of pure $N(0, 1)$ samples, which accounts for extreme-value behavior in multiple testing.

4 Results

4.1 Benjamini-Hochberg

Figure [4.1] demonstrates how the procedure works at $\alpha = 0.05$. Points below the red threshold line represent rejected null hypotheses. Points below the red threshold line are rejected null hypotheses. The high density of blue points (true positives) among them shows that the procedure selects many true discoveries early, which increases statistical power.

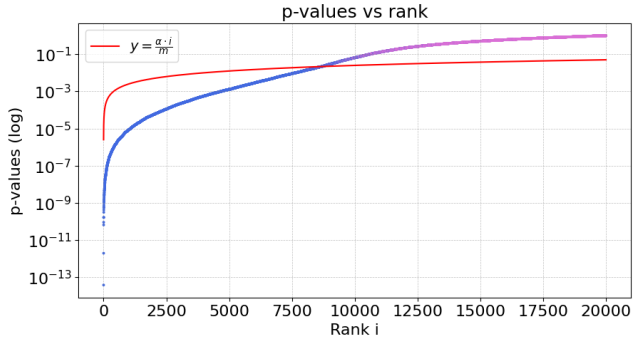

Figure 4.1: B-H p-values

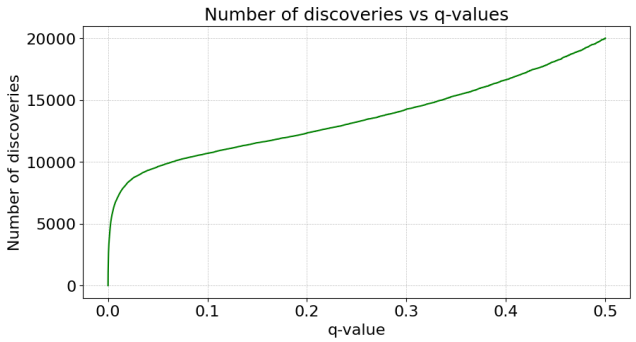

Figure 4.2: B-H q-values

The results become more clearly interpretable when expressed as q-values. Q-values represent the minimum FDR level at which a hypothesis would be called significant. Unlike p-values, q-values provide a more intuitive interpretation in the context of FDR. The q-value for the i-th ordered p-value is calculated as:

$$q_{(i)} = \min_{j \geq i} \left\{ \frac{m \cdot p_{(j)} \cdot \pi_0}{j} \right\}, \tag{5}$$

where π_0 represents the proportion of true null hypotheses.

The conventional BH algorithm assumes that all hypotheses under consideration are null

($\pi_0 = 1$), i.e., 100% of hypotheses are actually null hypotheses. In practice, such an assumption tends to yield over-conservative estimates. To address such a constraint, an estimation of π_0 was also proposed by John D. Storey in his 2002 paper. So by adjusting for π_0 (the fraction of true nulls), the q-value approach can increase power whenever $\pi_0 < 1$.

Figure [4.2] reveals a critical aspect of statistical power in multiple testing: the relationship between stringency (q-value threshold) and discovery rate. At low thresholds, many true effects are detected, indicating that strict FDR control can still yield many discoveries. As the threshold increases, the rate of new discoveries slows down. This illustrates the basic trade-off between FDR control and discovery rate.

To confirm the method's theoretical claims, a ground truth analysis using the known labels was conducted in the simulation. For each ordered hypothesis i, the actual FDR was calculated as

$$\frac{\sum_{j=1}^{i} \mathbf{1}\left[\text{label}_j = 0\right]}{i}. \tag{6}$$

Applying a monotonicity correction to these empirical FDR values yields ground truth q-values. Comparing them to the BH-computed q-values shows they match, validating the theory as can be seen on Fig. [4.3]. At high ranks the empirical FDR approaches 0.5, which corresponds to the true null proportion in our simulation.

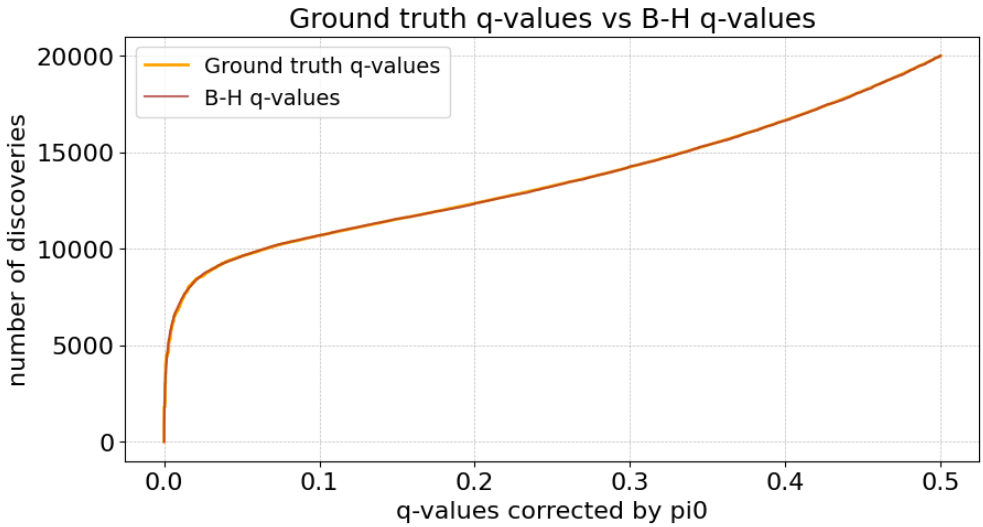

Figure 4.3: Ground truth q-values alignment

4.2 Target-decoy competition

Fig. [4.4] shows the implemented approach. FDR_1 consistently lies below the true FDR curve (olive), showing systematic underestimation. By undercounting false discoveries, FDR_1

becomes anti-conservative—at a given threshold it declares too many hits significant, risking excess false positives. In contrast, FDR_2 (salmon) always exceeds the true FDR, confirming its conservative bias. The built-in doubling of decoy counts appears to overcorrect, pushing the estimated FDR higher than the actual rate.

The black dashed line representing ground truth without decoys shows the theoretical limit, while the grey dashed line shows how decoy competition affects the empirical FDR calculation. Comparing the two ground truth curves reveals that including decoys (grey dashed) slightly raises the empirical FDR relative to ignoring them (black dashed). This indicates that the competition step itself introduces a small positive bias. Practically, at q-value thresholds between 0.05 and 0.10, FDR_1 underestimates by roughly 10–20% while FDR_2 overshoots by a similar margin. Seems that neither approach yields perfectly calibrated FDR control.

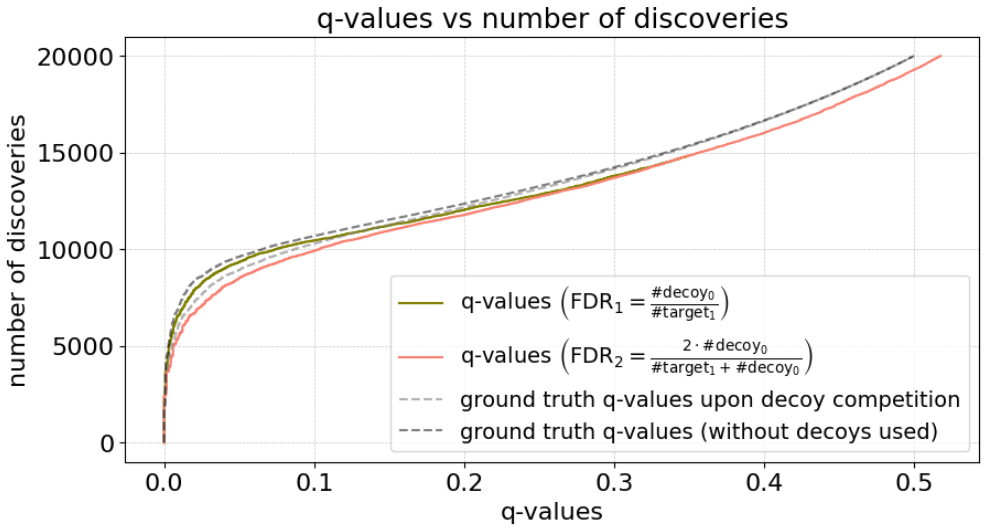

Figure 4.4: Target-decoy competition

Multiple target-decoy competition was also tested. In this approach the target–decoy competition is extended to include multiple decoy datasets, testing 1, 2, 3, 5, 10, 20, and 40 decoys across four different overlap scenarios.

Contrary to theoretical expectations adding more decoys does not reliably improve accuracy. In every overlap scenario from Fig. [4.5], the bias and variability of FDR estimates often increase rather than decrease as the number of decoys is raised.

Even when target and decoy distributions barely overlap, where distinguishing true signals should be easiest, raising decoy counts from 1 to higher values systematically widens the gap between FDR estimates and the ground truth. This indicates that multiple decoys can amplify systematic bias.

The expected stabilization is not observed from variance reduction. Increasing from one to

many decoys pushes the FDR curves farther right. More decoys thus increase the conservative bias instead of reducing it. As target and null distributions overlap more, the rightward bias grows and the curves become erratic. In moderate and high overlap settings, extra decoys make estimates both more conservative and less stable.

This systematic overestimation protects against false positives but also cuts true discoveries.

4.3 STDS, STDS-PIT, T-TDC, C-TDC, MIX-MAX

Figure [4.6] plots the log-ratio of actual FDR to estimated FDR versus the nominal threshold α for six methods, using a simulated data set of 500 spectra. A log-ratio > 0 means the actual false-discovery rate exceeds the estimate (the method is anti-conservative), while < 0 means the method is conservative.

STDS assumes all spectra lack true matches ($\pi_0 = 1$). When this is not true—as in our simulation where $\pi_0 = 0.5$—the method overestimates FDR. This happens because STDS treats all decoy hits as if they come from foreign spectra, overlooking that native spectra have lower false positive rates due to true matches in the database. Accurate FDR estimation with STDS depends on proper score calibration. If target and decoy scores are not directly comparable, the estimates can be biased due to distributional mismatches. STDS (blue) is strongly conservative: its estimated FDR is higher than the true FDR (log-ratio ~ -0.3 to -0.4 at higher α).

As noted by Noble and colleagues, STDS-PIT has a conceptual flaw. It estimates π_0 using p-values that test whether a spectrum is foreign, not whether the identification is incorrect. This matters because native spectra can still yield false matches if a random score exceeds the true one. This misinterpretation causes STDS-PIT to underestimate the true FDR. While it correctly adjusts for the presence of native spectra, it fails to account for incorrect matches among them, making the method anti-conservative. Despite this issue, STDS-PIT can still perform well in practice, especially when most spectra are foreign or when target and decoy score distributions are well separated. The log-ratio plot shows STDS-PIT (green line) maintaining a slightly positive log ratio across alpha values, indicating an anti-conservative FDR estimation. STDS-PIT showed nearly identical power to the theoretically optimal Benjamini-Hochberg procedure. But BH procedure (purple) shows a large positive bias: actual FDR is up to 50˘80% higher than nominal at low α, meaning BH under-estimates FDR and fails to control it accurately here.

C-TDC gives asymptotically unbiased FDR estimates when applied to the full list of winning targets and decoys. The factor of 2 accounts for the fact that decoy wins only capture one half of all false positives under the assumption of equal probability. In practice, researchers are usually

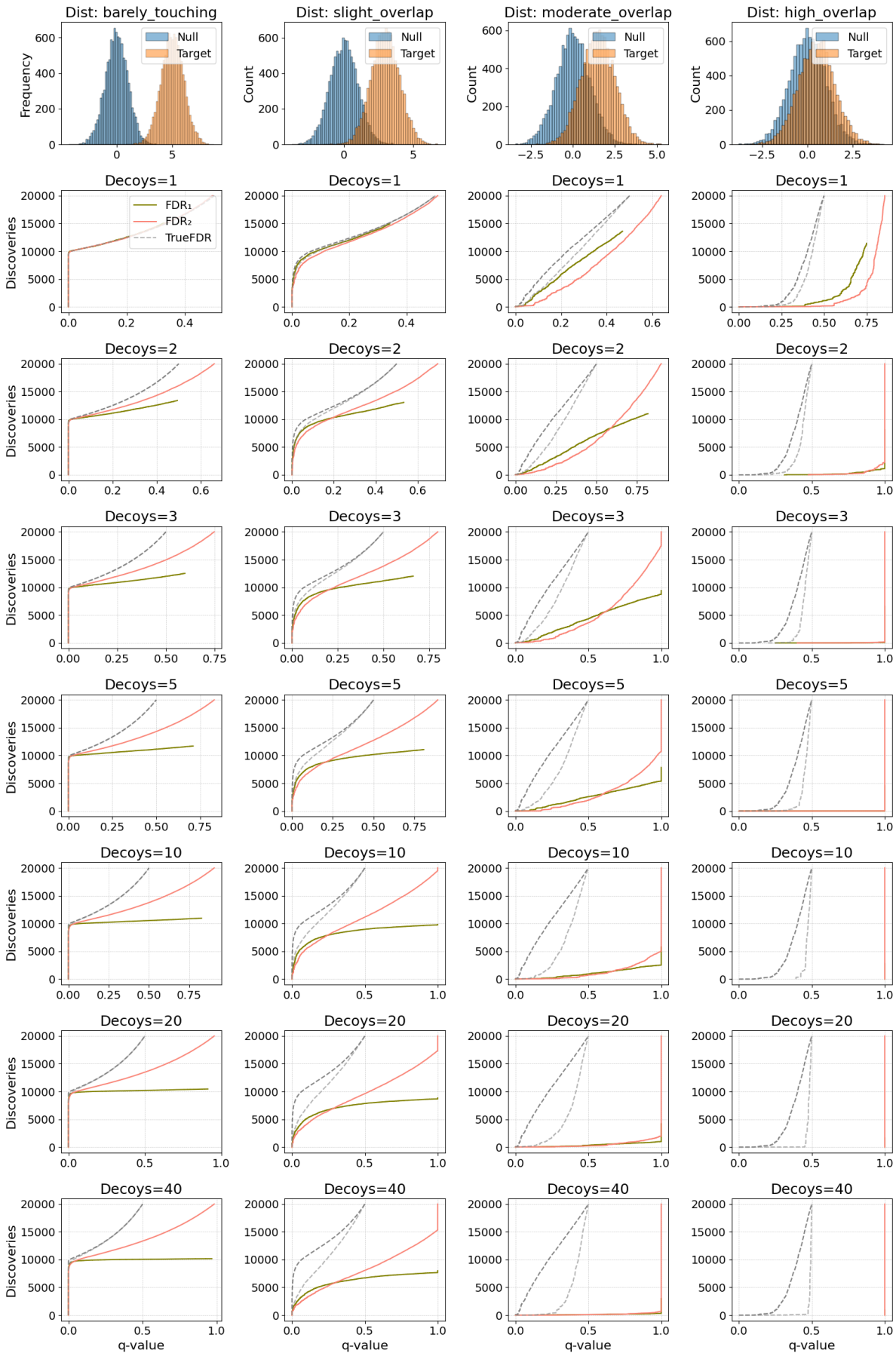

Figure 4.5: Target-decoy competition with multiple competition

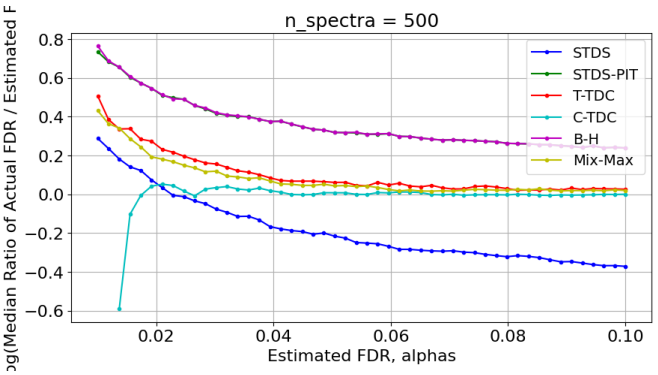

Figure 4.6: Log ratio of actual FDR to Estimated FDR

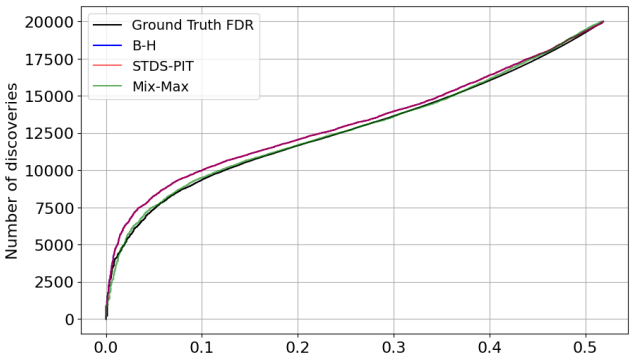

Figure 4.7: q-values vs discoveries

interested only in the target identifications. Since C-TDC estimates FDR over the combined list, using its results to evaluate target-only discoveries introduces a mismatch between the estimated false discovery rate and the list it's applied to. When applied to a target-only list, C-TDC becomes conservative. It includes decoy wins in the denominator while estimating false positives among a smaller number of targets, which leads to systematic overestimation of the FDR.

T-TDC doesn't depend on assumptions about score calibration across databases or about the proportion of foreign spectra. The competition itself calibrates the results. However some true matches can be missed because their corresponding decoy happened to score higher. This introduces some randomness into which correct identifications are kept, and can reduce power. But crucially, this filtering does not bias the FDR estimate for the targets that remain. Results show that T-TDC provides asymptotically unbiased FDR estimates as the number of spectra increases. In large datasets, this means the method becomes increasingly accurate.

Both C-TDC and T-TDC assume that each decoy match is statistically exchangeable with false target matches (i.e. decoy scores have the same null distribution as incorrect target scores). Under this assumption, C-TDC and T-TDC are asymptotically accurate for large data sets.

Mix-Max achieves more accurate FDR estimation by modeling different sources of false discoveries separately. This leads to more accurate estimates across many thresholds. As a result, it avoids being too strict or too lenient, which helps maintain higher statistical power. The mix-max method (yellow) is nearly unbiased: its log-ratio is around zero for most α, with only a mild positive bias at very low thresholds.

4.4 Cascaded search

Performance varies with the chosen FDR threshold. At a strict $\alpha = 0.01$, the cascade search finds 5,730 targets (5,700 true positives, 30 false positives), while the maximum dataset

finds 6,189 (6,029 true positives, 160 false positives). When α increases, the cascade gains more ground. At $\alpha = 0.05$, it identifies 10,748 targets versus 9,387 for the maximum approach. This advantage grows at further α. At $\alpha = 0.20$, cascade search locates 17,134 targets compared to 13,419 for the maximum method, demonstrating its superior discovery power under more liberal FDR control.

Fig. [4.8] shows the number of true positives (green) and false positives (red) in each dataset. Datasets processed earlier have more discoveries because they contain stronger signals. Later datasets yield fewer discoveries. This pattern supports the idea that analyzing data in order of decreasing signal density improves statistical power.

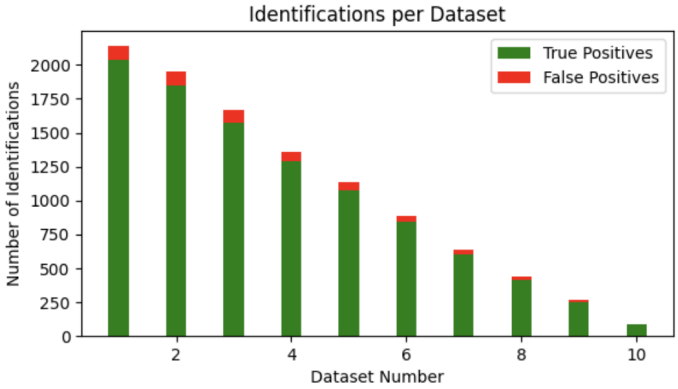

Figure 4.8: Stages of cascaded search

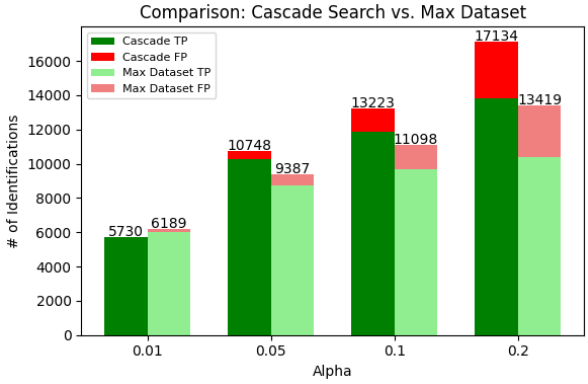

Figure 4.9: Comparison with max-dataset

5 Conclusion

This study evaluated the accuracy and power of several false discovery rate (FDR) estimation methods using simulation. Classical theoretical procedures were compared with empirical target-decoy methods, and more advanced approaches to understand trade-offs between statistical power and FDR control.

The results show clear performance differences. The Benjamini-Hochberg method, though theoretically valid under ideal assumptions, underestimated the FDR in our setting. This highlights the need to check assumptions in real applications. Procedures based on separated target/decoy searches (STDS and STDS-PIT) showed systematic biases: STDS consistently overestimates the true FDR while STDS-PIT underestimates it. In contrast, concatenated-DB methods such as the combined TDC and the T-TDC were essentially unbiased estimators of the FDR. The mix-max procedure, which combines target and decoy information under the assumption of well-calibrated scores, likewise produced accurate FDR estimates.

Clear trade-offs between conservativeness and discovery yield can be observed. Conservative

methods reported fewer identifications, whereas liberal methods reported more discoveries at the cost of inflated error. For example, STDS produced far fewer discoveries than Mix-Max, reflecting its conservative bias, whereas STDS-PIT reported more discoveries but underestimated the FDR. In contrast, cascade search dramatically improved statistical power, identifying substantially more true discoveries while maintaining nominal FDR control.

Thus, when accurate FDR estimation is critical, approaches like Mix-Max and T-TDC provide the most reliable performance. If the primary goal is to maximize the number of true discoveries, cascaded search strategies yield substantial improvements in power while still maintaining nominal FDR control. The choice between conservative and anti-conservative methods should be made based on the relative costs of false positives versus missed discoveries in the specific application context.

References

[1] Yoav Benjamini and Yosef Hochberg. "Controlling The False Discovery Rate - A Practical And Powerful Approach To Multiple Testing". In: *J. Royal Statist. Soc., Series B* 57 (Nov. 1995), pp. 289–300.

[2] Joshua Elias and Steven Gygi. "Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry". In: *Nature methods* 4 (Apr. 2007), pp. 207–14.

[3] Lukas Käll, John Storey, and William Noble. "Assigning Significance to Peptides Identified by Tandem Mass Spectrometry Using Decoy Databases". In: *Journal of proteome research* 7 (Feb. 2008), pp. 29–34.

[4] Uri Keich, Attila Kertesz-Farkas, and William Noble. "Improved False Discovery Rate Estimation Procedure for Shotgun Proteomics". In: *Journal of proteome research* 14 (July 2015).

[5] Attila Kertesz-Farkas, Uri Keich, and William Noble. "Tandem Mass Spectrum Identification via Cascaded Search". In: *Journal of proteome research* 14 (June 2015). DOI: 10.1021/pr501173s.

[6] John Storey. "A Direct Approach to False Discovery Rates". In: *Journal of the Royal Statistical Society Series B* 64 (Aug. 2002), pp. 479–498.